DOCUMENT RESUME

ED 094 445 CS 500 815

TITLE	Status Report on Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, January 1-June 30, 1974. Report No. SR-37/38(1974).
INSTITUTION	Haskins Labs., New Haven, Conn.
REPORT NO	SR-37/38(1974)
PUB DATE	Jun 74
NOTE	273p.
EDRS PRICE	MF-$0.75 HC-$12.60 PLUS POSTAGE
DESCRIPTORS	Articulation (Speech); *Conference Reports; *Educational Research; Higher Education; *Language Development; Language Skills; Listening Skills; Psychomotor Skills; *Speech; Speech Skills; *Verbal Communication

ABSTRACT

This report, covering the period of January 1 to June 30, 1974, is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Among the 17 manuscripts and extended reports are "The Role of Speech in Language: Introduction to the Conference," "The Human Aspect of Speech," "From Continuous Signal to Discrete Message: Syllable to Phoneme," "The Evolution of Speech and Language," " Phonetic Feature Analyzers and the Processing of Speech in Infants," "An Experimental Evaluation of the EMG Data Processing System: Time Constant Choice for Digital Integration," "More on the Motor Organization of Speech Gestures," "Electromyographic Study of the Velum During Speech," "The Function of the Posterior Cricoarytenoid in Speech Articulation," "Laryngeal Activity Accompanying the Moment of Stuttering: A Preliminary Report of EMG Investigations," "Hemispheric Lateralization for Speech Perception in Stutterers," "Categories and Boundaries in Speech and Music," and "A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech." (RB)

ED 094445

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SR-37/38 (1974)

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 January - 30 June 1974

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

CS 500 815

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

National Science Foundation
Grant GS-28354

Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration
Contract V101(134)P-71

Office of Naval Research, Information Systems Branch
Contract N00014-67-A-0129-0001

Advanced Research Projects Agency, Information Processing Technology Office, under contract with the Office of Naval Research, Information Systems Branch
Contract N00014-67-A-0129-0002

National Institute of Child Health and Human Development
Contract NIH-71-2420

National Institutes of Health
General Research Support Grant RR-5596

HASKINS LABORATORIES

Personnel in Speech Research

Franklin S. Cooper, President and Research Director
Alvin M. Liberman,* Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson[1]
Fredericka Bell-Berti*
Gloria J. Borden*
Rene Collier[2]
James E. Cutting*
Ruth S. Day*
Michael F. Dorman*
Jane H. Gaitenby
Thomas J. Gay*
Katherine S. Harris*
Hajime Hirose[3]
Frances Ingemann[4]
Philip Lieberman*
Leigh Lisker*
Ignatius G. Mattingly*
Paul Mermelstein
Rose Nash[5]
Patrick W. Nye
Lawrence J. Raphael*
Donald P. Shankweiler*
George N. Sholes
Michael Studdert-Kennedy*
Michael T. Turvey[6]
Tatsujiro Ushijima[3]

Technical and Support Staff

Eric L. Andreasson
Louis W. G. Barton
Elizabeth P. Clark
Janeanne F. Gent
Donald S. Hailey
Diane Kewley-Port*
Sabina D. Koroluk
Christina R. LaColla
Roderick M. McGuire
Agnes M. McKeon
Susan C. Polgar*
William P. Scully
Richard S. Sharkany
Edward R. Wiley
David Zeichner

Students

James Bartlett**
Mark J. Blechner*
Susan Brady*
Nina de Jongh**
Susan Lea Donald*
G. Campbell Ellison*
Cathy L. Felix*
F. William Fischer**
Christopher F. Foard*
Carol A. Fowler*
Frances J. Freeman*
Gary M. Kuhn*
Andrea G. Levitt*
Terrance M. Nearey*
Barbara R. Pick*
Barbara Pober**
Timothy C. Rand*
Philip E. Rubin*
Helen Simon**
Elaine E. Thompson*

---

*Part-time
**Volunteer
[1] On leave-of-absence to Ramkhamhaeng University, Bangkok, Thailand.
[2] Visiting from University of Louvain, Belgium.
[3] Visiting from University of Tokyo, Japan.
[4] Visiting from University of Kansas, Lawrence.
[5] Visiting from Inter American University of Puerto Rico.
[6] On leave-of-absence to University of Sussex, Brighton, England.

CONTENTS

I. Manuscripts and Extended Reports

The Role of Speech in Language: Introduction to the Conference -- Alvin M. Liberman. 1

The Human Aspect of Speech -- Ignatius G. Mattingly. 5

From Continuous Signal to Discrete Message: Syllable to Phoneme -- Michael Studdert-Kennedy 13

The Evolution of Speech and Language -- Philip Lieberman 25

Phonetic Feature Analyzers and the Processing of Speech in Infants -- James E. Cutting and Peter D. Eimas. 45

An Experimental Evaluation of the EMG Data Processing System: Time Constant Choice for Digital Integration -- Diane Kewley-Port. 65

More on the Motor Organization of Speech Gestures -- Fredericka Bell-Berti and Katherine S. Harris. 73

Electromyographic Study of the Velum During Speech -- T. Ushijima and H. Hirose. 79

The Function of the Posterior Cricoarytenoid in Speech Articulation -- Hajime Hirose and Tatsujiro Ushijima 99

Laryngeal Activity Accompanying the Moment of Stuttering: A Preliminary Report of EMG Investigations -- Frances J. Freeman and Tatsujiro Ushijima. . 109

Hemispheric Lateralization for Speech Perception in Stutterers -- M. F. Dorman and R. J. Porter, Jr. 117

Dichotic Release from Masking: Further Results from Studies with Synthetic Speech Stimuli -- P. W. Nye, T. M. Nearey, and T. C. Rand. 123

Binaural Subjective Tones and Melodies Without Monaural Familiarity Cues -- Michael Kubovy, James E. Cutting, and Roderick McI. McGuire. 139

Categories and Boundaries in Speech and Music -- James E. Cutting and Burton S. Rosner 145

Different Speech-Processing Mechanisms Can be Reflected in the Results of Discrimination and Dichotic Listening Tasks -- James E. Cutting. 159

The Intelligibility of Synthetic Monosyllabic Words in Short, Syntactically Normal Sentences -- P. W. Nye and J. H. Gaitenby 169

A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech -- Paul Mermelstein 191

What Information Enables a Listener to Map a Talker's Vowel Space? -- Robert Verbrugge, Winifred Strange, and Donald Shankweiler 199

Consonant Environment Specifies Vowel Identity -- Winifred Strange, Robert Verbrugge, and Donald Shankweiler 209

Identification of Vowel Order: Concatenated Versus Formant-Connected Sequences -- M. F. Dorman, James E. Cutting, and Lawrence J. Raphael 217

On "Explaining" Vowel Duration Variation -- Leigh Lisker 225

Two Processes in Vowel Recognition: Inferences from Studies of Backward Masking -- M. F. Dorman, D. Kewley-Port, S. Brady, and M. T. Turvey. 233

Vowel and Nasal Durations in Vowel-Nasal-Consonant Sequences in American English: Spectrographic Studies -- Lawrence J. Raphael, Michael F. Dorman, Charles Tobin, and Frances Freeman 255

Vowel and Nasal Durations as Perceptual Cues to Voicing in Word-Final Stop Consonants -- Michael F. Dorman, Lawrence J. Raphael, Frances Freeman, and Charles Tobin. 263

II. Publications and Reports 273

III. Appendix: DDC and ERIC numbers (SR-21/22 - SR-35/36). 279

I. MANUSCRIPTS AND EXTENDED REPORTS

The Role of Speech in Language: Introduction to the Conference*

Alvin M. Liberman
Haskins Laboratories, New Haven, Conn.

Our topic--the role of speech in language--is not an established one: no one has made it the direct and primary object of his research. It is the more appropriate, therefore, that one of the chairmen of this conference describe what we had in mind as we made our plans.

Our point of departure was a question: do we increase our understanding of language when we take into account that it is spoken? Obviously, we chairmen would answer in the affirmative; we hope that you will want to answer similarly. If so, the aim of our conference will be to consider, not whether our understanding of language is increased, but how; you will want to count the ways. It is not my place, however, to answer for you, or to try to bias the direction your discussion will take. I should only say why we think the question is a reasonable one and likely to trigger a productive discussion.

Our belief in an organic connection between speech and language comes, improbably, from some loose assumptions about grammar and the fit of grammatical form to grammatical function. We find nothing wrong with the linguists' assertion that the function of grammar is to connect sound to meaning, except that it does not take us as far as we want to go. Why is the grammar complex, and what purpose, if any, is served by the forms these complications take? What gain might have served, in evolution, to select for the physiological processes that underlie man's use of these peculiarly complex grammatical codes?

It may be useful, first, to replace the words "sound" and "meaning" with the structures to which they are presumably related. Sound is no problem for us; it is produced by the vocal tract and received by the ear. These structures, the vocal tract and the ear, are one terminal of the connection that grammar makes. Let us call it, for want of a better term, the transmission terminal. What, then, is the other terminal? What structure serves similarly for meaning, and what shall we call it? Mincing no words, we shall suppose that the other terminal is a nonlinguistic intellect, the place where our cognitive apparatus is housed. We need not speculate about the nature of that intellect or its associated machinery, except to emphasize that, as we have already said, it is not linguistic, and that communication within it is carried out, not in linguistic terms, nor indeed in

---

*Paper presented at a National Institute of Child Health and Human Development conference, "The Role of Speech in Language," held at Columbia, Md., October 1973, under the chairmanship of Alvin M. Liberman and James F. Kavanagh; to be published in the conference proceedings, ed. by J. F. Kavanagh and J. E. Cutting (Cambridge, Mass.: MIT Press).

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

terms of such other special processes as vision or audition, but in some amodal code. Fodor, Bever, and Garrett (1974), whose view of this whole process is similar to the one presented here, have called this code "mentalese." In any case, we have at the one end a source--the intellect--from which linguistic messages originate and to which they are delivered, and at the other end a transmission terminal--the vocal tract for producing sounds and the ear for receiving them. Grammar is the code that connects the source to the transmitter, and we are back once again to the question: what is the function of a complex grammatical connection?

To see the function of a grammatical connection, it is helpful to consider what communication would be like without it. In such agrammatic communication each message would be represented straightforwardly by a signal. The rule governing the flow of information would describe a one-to-one relation between a list of all possible messages and a corresponding list of signals. Moving back and forth between source and transmitter, the information would be converted--for example, from a neural representation to an acoustic one--but not in any way restructured. Now from a logical point of view, there is nothing wrong with that kind of communication; such a simple cipher can, in principle, do everything that complexly encoded language does. But from a biological point of view it works well only if there is reasonable agreement in number between the potential messages and the distinctively different signals that can be efficiently produced and perceived. And there is the rub. Though we don't know exactly how many uniquely different sounds we can cope with, the number is surely quite small. It is even harder, of course, to estimate the number of potential messages; if it is to include all that our stored experience and cognitive machinery can generate, however, it must be enormously large. It follows, then, that if we are to communicate with an agrammatic system, the terminal transmission apparatus--the vocal tract and the ear--will set the limit on what we can say, and a very low limit it will be.

To say, as we do, that the number of messages is vastly greater than the number of useable signals is to suggest that the intellect and the transmission terminal are not well matched to each other. But there is no reason to suppose that they should be. These structures developed in evolution long before the appearance of language and in connection with biological functions--thinking and remembering, in the one case, and eating, breathing, and hearing, in the other--that had nothing to do with language or, to any considerable extent, with each other. So long as vocal communication was agrammatic, the number of messages that a presumably rich intellect could send was severely limited by the number of distinctively different signals that a poor transmission system could cope with. We see then that a function of grammatical codes is to restructure information to make it differentially appropriate for processing and long-term storage in an intellect, on the one hand, and for transmission through a vocal tract and an ear, on the other. If so, then, as Mattingly (1972) has suggested, grammatical processes may have evolved as a kind of interface, matching the potentialities of an intellect to the limitations of our devices for producing and perceiving sounds, and thus increasing vastly the efficiency with which we can communicate ideas.

It is possible, of course, that other important changes might also have occurred in the evolution of language. Thus, the nonlinguistic structures that grammar connects might themselves have been modified in the direction of reducing the mismatch. Indeed, in the case of the vocal tract, at the one end of the system, such modifications did occur: our vocal tract differs anatomically from other primate vocal tracts, and in ways that appear to make it possible for us to produce a greater variety of sounds. If we had to speak with the vocal tract of an ape, the grammatical interface would have a larger matching job to do, and would presumably be that much more complex. As to what might have happened to the intellect, at the other end of the system, we hardly know how to ask the question; we can only speculate that our intellectual processes might in some unspecified way be better fitted to language than those of our nonhuman relatives. But such considerations, important though they may be, do not require an essential modification in our assumption about grammar and its function; we may still believe that the function of grammar is to reshape information so as to make it differentially appropriate for an intellect and a transmission system.

We come now to the basis for our assumption that speech is an organic part of language. Taking grammar as the most distinctive characteristic of language, and assuming that it evolved as a matching interface, we find it reasonable to suppose that its form would somehow reflect the characteristics of the nonlinguistic structures--intellect and transmitter--that it connects. Just how the grammar reflects those characteristics, and how the strength of the reflection varies with the distance from either of the nonlinguistic terminals, must be an empirical question. But if our view of the function of grammar and of its evolution is at all correct, then we should suppose that important aspects of language are as they are because language is normally spoken and heard.

To look at grammatical processes from that functional point of view, we could begin either at the intellectual end and work downward, or at the speech end and work upward. Beginning at the intellectual end has its attractions: we are closer there to the semantic and syntactic activities that have traditionally been thought of as the essence of language. But starting at the speech end, which is what we propose to do in this conference, does not necessarily lead us away from the distinctive characteristics of language, and it has the great advantage that the processes we will be concerned with are more readily available to scientific investigation. Taking speech to comprise the part of language that extends from the phonetic (or more abstractly phonological) message to the sound, we can frame our questions quite pointedly and reasonably hope to find some interesting answers. Thus, we can ask about the shape of the phonetic message and wonder, with some hope of satisfying our curiosity, whether there is anything like it in nonhuman communication? We can ask, further, what is required of the vocal tract and the ear if the phonetic message is to be efficiently communicated? Can these requirements be met straightforwardly, given the characteristics of the vocal tract and the ear, or is there a need for grammatical interfacing--a kind of speech grammar--even at this first, lowest stage of the system? If there is such a grammatical interface, what is its form and how well does the form fit the function? In the evolution of this system, did the auditory components of the transmission system change, as the vocal tract apparently did; if so, did the changes reduce the mismatch with the requirements of phonetic communication, thus making the grammatical interface less complicated than it otherwise would have been? Are speech production and perception unique to man, and if so, what are their unique attributes? What are the conditions for the development of speech in the human infant, what is its time course, and what evidence, if any, do we find there for a species-specific, innate predisposition to language? What happens to grammar when human beings who have normal intellectual apparatus must interface, not to the vocal tract and ear, but, as in the case of deaf mutes, to visible gestures? What can we learn, in other words, about the function of grammar by studying sign language? More generally, what can we say of the role of phonology in language? Looking at all of grammar, what evidence do we find of accommodation to the limitations of the vocal tract and the ear, and what formal resemblances, if any, do we see among syntax, phonology, and speech?

But these questions are only examples. You may or may not want to deal with some of them. At all events, we expect that you will ask better ones. Our aim has been only to illustrate a functional approach to language and grammar that we hope you will want to take as you explore the role of speech in language.

REFERENCES

Fodor, J., T. Bever, and M. Garrett. (1974) The Psychology of Language. (New York: McGraw-Hill).
Mattingly, I. G. (1972) Speech cues of sign stimuli. Amer. Scient. 60, 327-337.

The Human Aspect of Speech*

Ignatius G. Mattingly[+]
Haskins Laboratories, New Haven, Conn.

Underlying the recent work of many of us at this conference is the question I would like to consider today. It may at first seem strange and unnecessary, but I hope to persuade you of its relevance and importance. My question is, what aspects of speech, if any, are peculiarly and distinctively human?

At one time, it would have been enough to answer that speech is the vehicle for language, and only human beings have language. Both of those propositions are now in dispute. Some people seriously question whether there is really any justification for reserving the term "language" for human communication: Premack (1972) and Lieberman (1973) suggest that chimpanzees have language. Students of sign and gesture point out that early man may have communicated linguistically without benefit of speech (Hewes, 1973), and that many deaf persons certainly do so (Bellugi and Fischer, 1972; Stokoe, this conference); speech is perhaps only one of several vehicles for language. Moreover, there is reason to believe that speech evolved independently of language: structural parallels have been noted between speech and various animal communication systems that no one would call languages (Mattingly, 1972). Thus we should not attach undue significance to the fact that speech is specific to man; its peculiarities might prove on closer observation to be of a not very profound kind, like the details that distinguish the courtship display of one species of gull from that of another (Tinbergen, 1951). But in the face of these considerations, I would maintain that there is a truly human aspect of speech, something that marks it unmistakably as a product of man's cognitive powers.

Let me begin my pursuit of the human aspect of speech with a very general account of linguistic capacity. We suppose, with the generative grammarians, that the speaker-hearer has to deal with two significant versions of an utterance: a phonetic representation and a semantic representation (Chomsky, 1965). The phonetic representation of the utterance is in a form suitable for transmission by the vocal apparatus; the semantic representation is in a form suitable for storage in long-term memory (Liberman, Mattingly, and Turvey, 1972). It is convenient to conceive of both of these representations as _n_-dimensional arrays of features. The features of the phonetic representation are few in number and refer to the properties of the vocal tract (Chomsky and Halle, 1968), while those of the semantic representation are presumably far more numerous and refer to the whole of human experience. In the course of speaking or understanding an utterance, the speaker-hearer forms both of these representations. He can do so because he knows, tacitly, how the phonetic representation relates to the acoustic speech signal, how the semantic representation relates to the contents of long-term memory, and how the two representations relate to one another. This way of describing linguistic capacity suggests that phonetics, grammar, and semantics exhibit significant parallels, and this is just the impression I am striving to create. Let me try to bring out the parallelism by looking at each of these forms of cognitive activity in turn.

---

*Paper presented at a National Institute of Child Health and Human Development conference, "The Role of Speech in Language," held at Columbia, Md., October 1973; to be published in the conference proceedings, ed. by J. F. Kavanagh and J. E. Cutting (Cambridge, Mass.: MIT Press).

[+] Also University of Connecticut, Storrs.

Speech differs in interesting ways from other natural communication systems. To be sure, some animal communication systems share many "design features" (Hockett and Altmann, 1968) with speech, and there are striking parallels between the perception of speech and the perception of "sign stimuli" (Mattingly, 1972). But for none of these systems is it difficult to imagine how the perceptual powers of the users can cope with the amount of information known to be contained in the signal. The messages are typically very simple indeed. They are central to the survival of individuals and species, but their information content is low. With speech, however, the problem is to explain how the system can convey as much information as it does, without overwhelming the ear. Most of us will recall Liberman's (in Kavanagh, 1968) account of the obstacles encountered in developing various alphabets composed of discrete sounds to be used in a reading machine for the blind. As in speech, linguistic information was being communicated by an acoustic signal. Yet none of the sound alphabets could be understood at information rates comparable to that of natural speech. At a considerably lower rate, the individual alphabet sounds merged in a buzz (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967).

Again, consider the difference between speech and various possible gesture systems that might conceivably convey linguistic information. The comparison is the more appropriate because speech itself is a system of articulatory gestures. Perhaps the simplest possible gesture system would be one consisting of gross bodily movement--shrugging one's shoulders, tossing one's head, turning one's back. Such gestures, we know, can convey attitudinal meaning either alone or in conjunction with speech (Eibl-Eibesfeldt, 1970). But obviously the repertoire is not large and the semantic possibilities are limited. However, if we allow independent movements of the arms, the potential repertoire of gestures increases. If the two hands work together, and can point to or touch the various parts of the upper body, a still larger repertoire is available, and the perceiver can concentrate his attention on a fairly small part of his visual field. This is the method of sign language. Finally, consider facial gestures. The various physiognomic features can move independently and quite rapidly, and they are collocated in a fairly small visual area. Nonhuman primates communicate with facial gestures and human beings use them expressively, exploiting the extreme mobility of the face to express an extraordinary range of attitudes. One could even conceive of a form of sign language in which linguistic information would be transmitted by rapid and quasi-independent motions of brow, eye, nose, mouth, and chin. Such a physiognomic communications system might in principle carry more information than a manual system. However, it would be difficult or impossible to track the concurrent movements of the various facial features. Perhaps the motion of two hands is as much as the eye can comfortably follow.

But speech apparently offers a way of overcoming this kind of limitation. This is in a way rather surprising, for speech would not seem to be a highly efficient communication system. In certain respects it resembles our imaginary physiognomic system: a group of collocated articulators--larynx, tongue, velum, jaw, lips--are moving quasi-independently, and the perceiver has the difficult task of following these different movements. But there is a further problem. The perceiver of physiognomic gestures has a direct and continuous display, whereas the display available to the perceiver of speech is partial and indirect: it consists of the sounds that happen to be made as air passes through the shifting cavities and passages formed by the moving articulators. Yet from the indirect record contained in the acoustic signal, the listener can extract without difficulty the information carried by the articulatory gestures.

Speech can circumvent the limitations on both ear and eye because, unlike human gesture systems and animal communication systems, it is "encoded" (Liberman et al., 1967). To clarify what this means, consider the mapping of the phonetic representation on the speech signal. Most of the information is carried by a few prominent acoustic features: the fundamental frequency, the first two or three formants, the plosive bursts, and the patches of fricative noise. The rest of the signal can be discarded with little or no loss in intelligibility. Moreover, the acoustic signal does not consist of temporal segments corresponding to successive phones. Rather, the acoustic features typically carry information about two or more phones in parallel. These economies mean that the load on the input channel is much lighter than it would be if the perceiver had to attend to separate acoustic units, or to a visible array of gestures. Thus, phonetic information can be transmitted at a much higher rate.

The price for this gain at the input is the complexity with which the cues that are the basic data for speech perception are represented by the acoustic features (Mattingly and Liberman, 1969). Thus we find that the cues for two or more successive phones may be carried simultaneously by the same acoustic event: a second-formant transition cues the place of articulation of a stop, and the place of the adjacent vowel. On the other hand, different cues for one phone may be far apart: when two vowels are separated by a consonant, the place of the second vowel may be cued not only by the quasi-steady-state of its second formant, and the adjacent transition, but also by the second-formant transition between the first vowel and the consonant. Finally, quite different cues will signal the same phone in different environments: an alveolar consonant is cued by a rising second-formant before a front vowel and a falling second-formant before a back vowel. Indeed, the dispersal of information in time and frequency offers ample justification for Hockett's comparison of the speech signal to smashed Easter eggs (Hockett, 1955:210).

When we investigate the sources of this complexity, we find that a restructuring of information occurs in at least three different ways. The most obvious restructuring is an acoustic one. Variation in the spectrum of the speech signal is determined in part by the changing shapes of the vocal tract cavities (Fant, 1960). The relation of cavity shape to spectrum is hardly straightforward and sometimes ambiguous, but it is the movements of the articulators that are significant, and these are only indirectly reflected in the changes in cavity shapes over time. However, the matter is even more complicated. If an articulator consistently moved in such a way that, as a consequence of each motion, it attained a target position associated with some phonetic value, it would be possible to relate to each phone a target articulatory configuration, and hence a target shape and a target spectrum. But while such targets can be hypothesized, they are actually attained only in simple cases. More commonly, targets are merely approached, and the different articulators participating in the production of one phone do not generally come closest to their targets at the same time (Lindblom, 1963). Furthermore, the motion of an articulator is ordinarily complex, determined by preceding and following phones as well as by the current one (Öhmann, 1966). With electromyographic techniques, the different gestures underlying complex articulatory motion can frequently be distinguished at the neuromotor level (K. S. Harris, personal communication) as commands from different muscles. Yet restructuring can take place even at this stage of production; the muscle commands themselves are sensitive to phonetic context (MacNeilage and DeClerk, 1969).

Thus the task of speech perception is even more complicated than we originally suggested. What the listener has to recover from the acoustic data is not the mere physical motion of the vocal organs but the articulatory plan that is realized in this motion. How can we do this? At least part of the answer is that he is able to bring to his task information that severely constrains his perceptual hypotheses. My colleagues at Haskins Laboratories have argued that the listener has tacit knowledge of certain properties of the vocal tract, and they have proposed a "motor theory of speech perception" (Liberman et al., 1967). Rather than reviewing their arguments, let us take it that the theory is essentially correct, and consider the kind of knowledge the theory imputes to the listener. Certainly, it is not the kind of knowledge that could be deduced from communication theory, or even from an analysis of the acoustic speech signal alone. The vocal tract is a highly eccentric collection of disparate structures with distinct primary functions. Though it has undergone some remodeling to make it a more serviceable signaling device (Lieberman, 1968), it is essentially a bizarre arrangement that can be rationalized only in evolutionary terms. Nor is it the kind of knowledge that the listener might be supposed to derive from his own experience as a speaker. Experience in speaking is neither necessary, as is known from clinical cases in which damage to, or congenital deformation of the vocal tract does not interfere with speech perception (Lenneberg, 1967), nor sufficient, since it would not be adequately generalized knowledge. We need to assume that the listener's tacit knowledge is of a more abstract character if we are to account for his ability to recover phonetic information from the output of vocal tracts of different shapes and sizes. We might imagine his knowledge as the equivalent of a dynamic vocal tract model, an ideal speech synthesizer.[1] With a few adjustments, the model is good for any speaker. It is a highly selective model, enabling the processes of perception to extract information from the signal received by the ear, even though this signal is a complexly encoded record of articulation.

If such a model seems over-elaborate, consider that it is required also to constrain the articulatory plan of an utterance if the speaker is not to make inconsistent or impossible demands on his articulators, and to monitor the utterance as it is being produced. Production and perception are regulated by the same tacit knowledge. The model is also needed to account for the fact that the infant must deduce the phonetic rules of his language from speech produced by adult vocal tracts very different from his own (Lieberman, Crelin, and Klatt, 1972), and must learn to manipulate his own vocal tract, accommodating to its individual variations and to its changes in shape and size as it matures.

---

[1] I do not mean by these comparisons to suggest a "process" model. Neither the proposed model nor any synthesizer actually recapitulates the processes of production; rather, they demonstrate the relationship of selected phonetic variables to acoustic output.

Speech perception, then, is a very powerful process because it applies to the analysis of a certain kind of very complex data a profound, specialized knowledge about such data. This cannot be said of sign language, or of the gestural communication systems imputed to early man.

Let us turn now, adopting the conceptual framework of generative grammar, to the relationship between the phonetic and semantic representations. If we compare these two representations for some utterance, we find that in the phonetic representation, the elementary propositions of the semantic representation ("deep structure") are internally reordered and combined with one another, that some lexical morphemes have been deleted or anaphorically replaced, and that other morphemes have been introduced, any one of them perhaps representing two or more semantic or syntactic elements. At morpheme boundaries, as well as within morphemes, there is extensive phonological revision: sounds have been inserted, deleted, changed in one or more distinctive features, or transposed with one another. In short, the phonetic representation is an encoding of the semantic representation (Mattingly and Liberman, 1969). The effect of the encoding is to make the syntax of the utterance as compact and the articulation as efficient as possible. The speaker-hearer's ability to produce appropriate phonetic representations, as well as his ability to reconstruct the semantic representation from the phonetic representation, is dependent upon his competence: his internalized knowledge of the grammatical rules of his language. This grammar is so highly specified that he is able to judge the grammaticality of any utterance and to recover its deep structure. The grammar is generative; that is, the grammatical analysis of an utterance is its derivation from deep structure according to rules. Infinitely many such utterances can be derived.

A generative grammar plays a role in the production and the understanding of sentences similar to the role played by the vocal tract model in speech production and perception. It embodies the knowledge that enables the encoding to be correctly imposed and removed. The parallel may not seem immediately obvious because our way of knowing grammatical facts is very different from our way of knowing phonetic facts. Phonetic activity is to some extent observable, but grammatical activity is not. On the other hand, we have considerable intuitional insight into grammar and little or none into phonetics. Grammar is customarily presented as a system of formal rules rather than as a neurophysiological model.

This formalism, however, brings out certain interesting restrictions: the division of the grammar into components with different types of rules, the nonoccurrence of certain transformational patterns, the type of context that must be stated in a phonological rule, and so forth (Chomsky, 1957). These restrictions are not functional; they reflect indirectly the idiosyncrasies of the underlying neurophysiological apparatus, though we cannot observe it directly as a flesh and blood reality, as we can the vocal tract. Moreover, such formalism is what we would expect of our articulatory model if it is to be the abstract thing we have suggested, independent of particular vocal tracts yet capturing their essential common features.

Our claim, then, is that the listener's grammatical analysis and his phonetic analysis are really quite similar forms of cognitive activity.[2] In each case, the listener brings to a complex array of data tacit knowledge sufficiently well-specified for him to determine the underlying structure of the array and to remove the encoding.

Consider finally the question, how does the speaker get the semantic representations that he encodes linguistically? This is of course a special case of a problem central to cognitive psychology: how is information stored in and recovered from memory? We know that the capacity of long-term memory must be enormous, but we do not know how we thread our way through the labyrinth so readily. Yet some inferences useful to our present purposes are prompted by the familiar phenomenon of paraphrase (Liberman et al., 1972). If someone is asked to recall a sentence he has previously heard, he responds, typically, with a sentence that probably differs semantically (and also grammatically and phonetically) from the original, though probably also remaining semantically consistent with it. In fact, we would find it strange if on a certain day A were to say to B, "I'm coming tomorrow," and B were to report this the following day saying, "A said I'm coming tomorrow," without indicating by appropriate intonation that he was quoting A directly. We would expect rather a paraphrased response that takes into consideration changes in the time, the speaker, and the world ("A said that he's coming today."). The phenomenon of paraphrase suggests that the semantic representations of the sentences one hears are not ordinarily preserved in memory as separate items. Indeed, this would be a most inefficient way of storing information, given the redundancy of human discourse. Rather, the semantic content of the sentence is somehow incorporated into a more general record of experience that is the basis of both paraphrases and newly created sentences. In fact, the two cannot logically be distinguished.

Bartlett (1932) has suggested that experience is represented in memory by "schemata." A schema is not a chronological record of previous perceptions, each separately preserved, but rather an integration of these perceptions into an "organized setting" relating to a particular sort of experience. A new perception is drastically influenced by the schemata, and also modifies the schemata themselves. The same, presumably, can be said of the semantic representation of a newly heard sentence. It does not seem likely that there can be any simple mapping of parts of the semantic representation onto parts of the schema. The schema is the product of a great many semantic representations and, of course, of percepts of other kinds. On the other hand, a single semantic representation may conceivably modify many different aspects of the schema, and the nature of the modification may depend on the state of the schema itself. If we had available some convenient visual display of a schema in memory--an enormous spectrogram, as it were--we would doubtless find it very difficult to identify unambiguously the correlates of a particular sentence. Memorial processes are neither directly observable nor intuitively accessible, but we suspect that the integration of a semantic representation is a kind of encoding. It is similar in principle to the other encodings we have discussed, though different both in the content of what is encoded and in its scale, for the schema is an encoding not of one but of many sentences.

---

[2] Lieberman (1973) thinks that the encodedness of speech is peculiarly human, just as I do, but that the cognitive ability underlying language differs only quantitatively from the "logical" ability underlying the conditioned responses of lower animals.

But if old sentences are lost, where do new sentences come from? The general affinity between perception and recall, and Bartlett's shrewd comment about remembering, are suggestive: "...the organism would say, if it were able to express itself: 'this and this and this must have occurred, in order that my present state should be what it is'" (Bartlett, 1932:202).

Putting this in the terms we have been using, to recover semantic information is to produce a semantic representation that if encoded would prove to be consistent with the current state of the schemata. Moreover, for independent linguistic reasons, we want the semantic representations underlying the sentences the speaker produces to be formally similar to the semantic representations of the sentences he hears: there is only one kind of deep structure. What is needed, therefore, both for storage and for recall, are rules that relate semantic representations to schemata: a sort of generative grammar of memory. These rules must reflect the nature of human experience that can be remembered. They must also reflect any purely nonfunctional properties of memory, attributable to its evolutionary history. And since we would not expect the processes of recall to vary markedly depending on the schemata of the individual, the rules must be abstract and general enough to transcend such individual differences. The storage and recovery of semantic information in memory is thus a further instance of the kind of cognitive operation that we have already observed in speech and language.

To recapitulate, I have tried to trace a cognitive pattern common to the processes of language, speech, and memory. In each of these processes information is thoroughly reorganized for functional reasons. The relationship of the original information to its reorganized form is complex: we have termed it encoded. Recovery of the information is accomplished with the help of a grammar which specifies the relationship between the unencoded and the encoded form of the information, and whose formal properties consequently reflect the nature of the encoding device. We cannot say that this cognitive pattern is unknown in lower animals: for example, the spider's knowledge of his web seems not dissimilar. But surely only in man is the pattern so highly developed and so diversely manifested. Of these manifestations, speech is of special interest. It is not as complex a system as language or memory, and it does not claim our attention so immediately when we reflect on the character of our knowledge. But it exemplifies nonetheless a thoroughly and peculiarly human kind of knowing.

REFERENCES

Bartlett, F. C. (1932) Remembering. (Cambridge: Cambridge University Press).

Bellugi, U. and S. Fischer. (1972) A comparison of sign language and spoken language. Cognition 1, 173-200.

Chomsky, N. (1957) Syntactic Structures. (The Hague: Mouton).

Chomsky, N. (1965) Aspects of the Theory of Syntax. (Cambridge, Mass.: MIT Press).

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper and Row).

Eibl-Eibesfeldt, I. (1970) Ethology. (New York: Holt, Rinehart & Winston).

Fant, C. G. M. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Hewes, G. (1973) Primate communication and the gestural origin of language. Curr. Anthropol. *14*, 5-24.

Hockett, C. F. (1955) *A Manual of Phonology*. Memoir 11, Internat. J. Ling. (Baltimore, Md.: Waverly Press).

Hockett, C. F. and S. A. Altmann. (1968) A note on design features. In *Animal Communication*, ed. by T. A. Sebeok. (Bloomington, Ind.: Indiana University Press).

Kavanagh, J. F., ed. (1968) *Communicating by Language: The Reading Process*. (Bethesda, Md.: National Institute of Child Health and Human Development).

Lenneberg, E. H. (1967) *Biological Foundations of Language*. (New York: Wiley).

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. *74*, 431-461.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In *Coding Processes in Human Memory*, ed. by A. W. Melton and E. Martin. (Washington, D.C.: V. H. Winston).

Lieberman, P. (1968) Primate vocalizations and human linguistic ability. J. Acoust. Soc. Amer. *44*, 1574-1584.

Lieberman, P. (1973) On the evolution of language: A unified view. Cognition *2*, 59-94.

Lieberman, P., E. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. *74*, 287-307.

Lindblom, B. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. *35*, 1773-1781.

MacNeilage, P. and J. DeClerk. (1969) On the motor control of articulation in CVC monosyllables. J. Acoust. Soc. Amer. *45*, 1217-1233.

Mattingly, I. G. (1972) Speech cues and sign stimuli. Amer. Scient. *60*, 327-337.

Mattingly, I. G. and A. M. Liberman. (1969) The speech code and the physiology of language. In *Information Processing in the Nervous System*, ed. by K. N. Leibovic. (New York: Springer Verlag).

Öhmann, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Amer. *39*, 151-168.

Premack, D. (1972) Language in chimpanzee. Science *172*, 808-822.

Tinbergen, N. (1951) *A Study of Instinct*. (Oxford: Clarendon Press).

From Continuous Signal to Discrete Message: Syllable to Phoneme*

Michael Studdert-Kennedy[+]
Haskins Laboratories, New Haven, Conn.

All speech is syllabic. All languages constrain syllabic structure, and base their constraints on the natural phonological contrast between consonant and vowel. Certainly, the human vocal tract is capable of producing an indefinite sequence of vowels, and even of certain consonants. But the choice of all languages has been to alternate consonants, or consonant clusters, and vowels. What is the communicative function of this choice? Mattingly (this conference) has sketched the broad outline. He has pointed to the complex encoding of discrete phonetic elements into a more-or-less continuous acoustic signal; he has shown how the code permits unusually rapid and efficient transfer of information; he has drawn striking formal analogies among phonetics, grammar, and memory. My purpose is to examine two aspects of the phonetic code in more detail: first, the structure and function of the syllable as reflected in perception; second, more briefly, the underlying cerebral mechanisms. Taken together, they suggest a natural system elegantly adapted to its role in linguistic communication.

Perceptual Structure and Function of the Syllable

Let me say, to begin with, that precise definition of consonant and vowel[1] is even more difficult than the definition of syllable. For, as Sweet (1877) observed, "the boundary between vowel and consonant, like that between the different kingdoms of nature, cannot be drawn with absolute definiteness" (p. 51). Most of what I have to say deals only with stop consonants and relatively sustained monophthongal vowels. Our knowledge of speech perception comes largely from study of these two phonetic classes and their combination in a simple CV syllable. But I believe this knowledge can furnish quite general insights.

---

*Paper presented at a National Institute of Child Health and Human Development conference, "The Role of Speech in Language," held at Columbia, Md., October 1973; to be published in the conference proceedings, ed. by J. F. Kavanagh and J. E. Cutting (Cambridge, Mass.: MIT Press).

[+] Also Queens College and the Graduate Center of the City University of New York.

[1] "Consonant" and "vowel" refer to elements in the phonetic message rather than to their correlates in the acoustic signal. I have used the terms for both meanings in what follows and trust that context will make clear which is intended.

Acknowledgment: I am grateful to Alvin Liberman, Ignatius Mattingly, and Donald Shankweiler for comments and criticism.

The syllable is an articulatory and acoustic integer. Greek grammarians gave us the word for it: sullabē, "a taking together," syllable. Roman grammarians gave us words for the elements that are taken together: littera vocalis, "the voiced letter," vowel; and littera consonans,"the letter that sounds with" the vowel, consonant. Into the syllable the speaker encodes, and from it the listener decodes, these discrete, phonetic segments. Among the evidence for the psychological reality of such segments is the corpus of spoonerisms gathered by Fromkin (1971). Here, one aspect of her data is of particular interest. Speakers may blunder by exchanging syllables for syllables, consonants for consonants, and vowels for vowels, but they never metathesize across classes. They may say, "I'll have a slice of [bost rif]," or even "of [rist bof]," but never "of [iost brf]." The syllable, itself a functional unit, is compounded of consonant and vowel, each fulfilling some syllabic function that forbids their metathesis.

Stetson (1951) gave us some understanding of these functions in production. He recognized the syllable as the fundamental unit of speech, the unit of stress contrast, of rhythm and meter. His motor definition of the syllable as a "chest pulse" has not stood up, at least for unstressed syllables (Ladefoged, 1967: Ch. 1). But his description of the time course of the syllable is still useful. He described the consonant-vowel-consonant (CVC) syllable as "a single ballistic movement," composed of release, nucleus, and arrest. Similarly, Pike (1943) described speech as alternate constrictions and openings of the vocal tract.

One obvious acoustic correlate of the syllabic movement can be seen in the amplitude display of a spectrogram. The onset of an open (CV) syllable tends to yield a low, but rapidly rising, amplitude, the nucleus a relatively sustained peak. For a longer utterance we can make a crude syllabic count from the amplitude peaks. The count is crude because many utterances (the word "tomorrow" [təmɑro], for example) may yield a single peak (due only in part to the sluggishness of the spectrograph's amplitude integrating response), even though we know they contain several syllables. Where amplitude fails to reveal syllabic structure, formant pattern may serve: the initial pattern tends to display a rapid movement, or scatter, of energy over the frequency domain, while the later portion tends to be relatively stable and sustained (cf. Malmberg, 1955). Taken together, changes in amplitude and frequency offer an acoustic contrast between the beginning and the end of a CV syllable, between its onset and its nucleus. The event is unitary, but its character changes as it occurs. This ill-defined acoustic contrast provides the auditory ground for the perceptual consonant and vowel.

Let us turn, first, to the phenomenon of categorical perception. Experiments have repeatedly revealed differences between stop consonants and vowels in their patterns of identification and discrimination. Stimuli for these experiments consist of a dozen or so synthetic speech sounds distributed in equal acoustic steps along a continuum ranging across two or more phonetic categories (from, say, [b] to [d] to [g] or from [i] to [I] to [ɛ]). In identification, listeners are asked to assign each of the stimuli, presented repeatedly and in random order, to one of the phonetic classes. They then assign consonants more consistently than they do vowels, particularly those tokens close to a boundary between phonetic classes. In other words, they identify consonants absolutely or "categorically," independently of the test context, while they identify vowels relatively or "continuously," with marked contextual effects (Liberman, Harris, Kinney, and Lane, 1961; Fry, Abramson, Eimas, and Liberman, 1962).

Here we have the first, and oldest, indication that listeners have a longer short-term auditory store for vowels than for consonants. Recently, Sawusch and Pisoni (1973) have elaborated with the finding that vowels, like nonspeech tones, are susceptible to psychophysical anchoring effects: the boundary between synthetic vowels along an acoustic continuum is shifted toward the vowel that occurs most frequently in the test. For consonants, the effect is absent: the listener relies on some internal standard, less readily subverted by test composition. Note, incidentally, that if a vowel, already assigned to its phonetic class, is to affect phonetic assignment of following vowels, the listener must retain an auditory image, or echo, of the vowel even after he has identified it. The process of identification does not therefore terminate auditory display: auditory store and phonetic store can exist simultaneously (cf. Wood, 1973b).

In the related discrimination task, typically administered in ABX format, a listener is called on to discriminate between pairs of tokens separated by one or more equal acoustic steps along the synthetic continuum. If these tokens are drawn from different phonetic classes, discriminative performance is high for both consonants and vowels. If they are drawn from the same phonetic class, discriminative performance drops slightly for vowels and considerably, to a point little better than chance, for consonants. In other words, a listener discriminates between vowels at a relatively high level whether he assigns them to different phonetic categories or not: his discrimination is more or less independent of identification, much as it is for nonspeech sounds (Mattingly, Liberman, Syrdal, and Halwes, 1971). But a listener can reliably discriminate between consonants only if he assigns them to different phonetic categories: his discrimination depends upon and, to a fair degree, can be predicted from his phonetic assignments (Liberman et al., 1961; Studdert-Kennedy, Liberman, Harris, and Cooper, 1970).

Early accounts of this phenomenon pointed to articulatory differences between consonants and vowels. But their acoustic differences have proved more crucial. Stevens (1968) remarked the brief, transient nature of consonantal acoustic cues, and Sachs (1969) showed that vowels were more categorically perceived if their duration and acoustic stability were reduced by CVC context. Lane (1965) pointed to the greater duration and intensity of the vowels and showed that they were more categorically perceived if they were "degraded" by being presented for discrimination in noise.

The role of auditory, or echoic, memory, implicit in the work of Stevens, of Sachs, and of Lane, was made explicit by Fujisaki and Kawashima (1969, 1970). They argued that the listener's poor auditory memory for consonants forced him to rely, for ABX discrimination, on phonetic memory. They formulated a mathematical model of the process and showed that, if they reduced vowel duration sufficiently, their model would predict quite accurately the discrimination of both consonants and vowels from listeners' phonetic identifications.

Pisoni (1971, 1973) made a direct test of this account. He varied the intratrial interval in an AX "same"-"different" task. For vowels, an increase in the A-X interval (with a presumed decrease in clarity of A's auditory store) led to a decrease in the likelihood that a listener would judge two acoustically different, but phonetically identical, tokens as "different." For consonants, there was no significant effect. In several other experiments, Pisoni (in press) has shown that the degree to which vowels are perceived categorically (measured by the degree to which phonetic class predicts discriminability) may be varied by manipulating the degree to which auditory memory is made available in the experimental task. Note, however, that while a fair degree of categorical perception of vowels can be readily induced, continuous perception of consonants is much more difficult (Pisoni and Lazarus, in press). The listener's auditory memory for consonants is intrinsically short.

Consider next dichotic ear advantages. As is well known, Kimura (1961a, 1961b, 1967) showed that if different digits are presented to opposite ears at the same time (i.e., dichotically), those presented to the right ear are recalled more accurately than those presented to the left ear. She attributed the effect to specialization of the left cerebral hemisphere for language functions and to stronger contralateral than ipsilateral ear-to-hemisphere connections. The effect and her interpretation have been repeatedly supported.

Shankweiler and Studdert-Kennedy (1967) used Kimura's technique to probe the processes of speech perception. They showed that the right-ear advantage did not depend on higher language processes, since it could be obtained with pairs of nonsense syllables differing only in an initial or final stop consonant. Furthermore, they showed that the effect did not appear if the competing syllables were steady-state vowels or CVC syllables differing only in their vowels (Studdert-Kennedy and Shankweiler, 1970). Following Kimura and others (Milner, Taylor, and Sperry, 1968; Sparks and Geschwind, 1968), they assumed that ipsilateral ear-to-hemisphere connections were inhibited by dichotic competition, so that, while right-ear inputs reached the left (language) hemisphere by a direct contralateral path, left-ear inputs traveled an indirect route, contralaterally to the right hemisphere, then laterally across the corpus callosum to the left hemisphere. The ear advantage was due to loss of auditory information from the left-ear signal as it traveled its indirect path to the language hemisphere. The consonant-vowel difference in ear advantage could then be attributed to the same acoustic factors as their differences in degree of categorical perception: the vowel portion of the signal, being more intense and of greater duration than the consonant, suffers less loss or "degradation" on the left-ear-to-left-hemisphere indirect path, and so yields no reliable right-ear advantage. This is probably not the whole story, since nonacoustic, attentional factors have also been implicated (e.g., Spellacy and Blumstein, 1970). However, Weiss and House (1973) have played for dichotic ear advantages the role that Lane (1965) played for categorical perception: they have shown that a right-ear advantage appears for vowels if the vowels are presented dichotically in noise.

Yet another experimental paradigm yielding stop consonant-vowel differences, this time directly in short-term memory, is due to Crowder (1971, 1972, 1973). If subjects are given, one item at a time, a span-length list of digits for immediate, ordered recall, they recall the last several digits more accurately when the list has been presented by ear than when it has been presented by eye. This modality difference presumably reflects the operation of separate modality stores, the short-term auditory store being more retentive than the visual (Crowder and Morton, 1969). This interpretation is supported by the fact that the advantage to the most recent auditory items (recency effect) is reduced or eliminated if the list ends with a redundant verbal item, as a signal for the subject to begin recall (suffix effect). That the suffix interferes with auditory store is suggested by the fact that the effect occurs for either backwards or forwards speech, is reduced if list and suffix are spoken in different voices, and is absent if the suffix is a tone.

The finding of interest in the present context is simply that while all three effects (modality, recency, suffix) are observed for lists of CV syllables differing in their vowels (e.g., random repetitions of /gæ , ga, gʌ/ or of /bi, bo, bu/), none of them is observed for CV or VC syllables differing in their initial or final stop consonant (random repetitions of /ba, da, ga/ or of /ab, ad, ag/). Crowder (1971) concludes that vowels are included in precategorical, auditory store, but that stop consonants are excluded. Liberman, Mattingly, and Turvey (1972) agree with this conclusion, arguing further that phonetic decoding of the stops "strips away all auditory information," that phonetic classification terminates auditory display.

While this interpretation is unlikely to be correct (cf. Wood, 1973b), the difficulty of retaining auditory information about stop consonants is again suggested by results from a fourth experimental paradigm, devised by Dorman (1973). He synthesized three three-formant sounds: a 250 msec /bæ /, a 250 msec /æ /, and a "chirp" consisting of the first 50 msec of the synthesized /bæ /. The "chirp" contained all the acoustic information necessary for phonetic classification of the initial /b/, but, separated from the following vowel, no longer sounded like speech. Dorman next varied the intensity of the "chirp" and of the first 50 msec of the two speech sounds in two steps: 0, -7.5, and -9 db. He then presented these stimuli in pairs, each member drawn from the same stimulus type, to ten subjects and asked them to judge whether the initial intensities of the pairs were the "same" or "different." The results were that every subject gave close to 100% performance on the vowels and "chirps," and close to 50% performance, or chance, on the CV syllables. In other words, asked to judge acoustic differences irrelevant to segmental classification, subjects could detect those differences in vowels or nonspeech sounds, but not in stop consonants.

We must not exaggerate. Many experiments have demonstrated that listeners do retain at least some "echo," however rapidly fading, of stop consonants. All studies of categorical perception reveal some margin of auditory discriminability within stop consonant categories, and several experimenters have tested this directly. Barclay (1972), for example, showed that listeners could reliably judge variants of /d/, drawn from a synthetic continuum, as more like /b/ or more like /g/. Pisoni and Tash (in press) found that reaction times for "same" responses were faster to pairs of acoustically identical stop-vowel syllables than to pairs of phonetically identical, but acoustically different, syllables.

Furthermore, Darwin and Baddeley (in press) have recently challenged Crowder's interpretation that stop consonants are excluded from precategorical store. They have shown that a moderate recency effect may be obtained with consonants if the syllables in the list are acoustically distinct (/ʃa, ma, ga/), and even more if the consonants are in syllable-final position. They argue that listeners cannot make use of their auditory store of the later items in a list if those items are acoustically similar and confusable, as are /ba, da, ga/. They support their argument by demonstrating that the recency effect can be reduced or abolished for vowels if the vowels are very brief (30 msec of steady-state in a 60 msec CV syllable) and occupy neighboring positions on an F1-F2 plot. They conclude that "the consonant-vowel distinction is largely irrelevant," and they propose "acoustic confusability" as the determining variable.

However, among the determinants of "acoustic confusability" are the very acoustic factors that distinguish stop consonants from vowels, namely energy and spectral stability. We have reviewed evidence from four experimental paradigms in which consonant and vowel perception differ. In three of these (and, no doubt, if one chooses, in the fourth) the differences can be reduced or eliminated by taxing the listener's auditory memory for the vowel or by sensitizing it for the consonant. But these qualifications do not mitigate the consonant-vowel differences: they merely emphasize that the differences are there to be eliminated. There is little question that consonants are less securely stored in auditory memory than vowels. [For further discussion see Studdert-Kennedy (in press-a, in press-b).]

Plausible communicative functions for these differences are not hard to find. Consider, first, vowel duration. Long duration is not necessary for recognition. We can identify a vowel quite accurately and very rapidly from little more than one or two glottal pulses, lasting 10 to 20 msec. Yet in running speech vowels last 10 to 20 times as long. The increased length may be segmentally redundant, but it permits the speaker to display other useful information: variations in fundamental frequency, duration, and intensity within and across vowels offer possible contrasts in stress and intonation, and increase the potential phonetic range (as in tone languages). Of course, these gains also reduce the rate at which segmental information can be transferred, increase the duration of auditory store, and open the vowel to contextual effects, the more so, the larger the phonetic repertoire. A language built on vowels, like a language of cries, would be limited and cumbersome.

Adding consonantal "attack" to the vowel inserts a segment of acoustic contrast between the vowels, reduces vowel context effects, and increases phonetic range. The attack, itself part of the vowel (the two produced by "a single ballistic movement"), is brief, and so increases the rate of information transfer. Despite its brevity, the attack has a pattern arrayed in time and the full duration of its trajectory into the vowel is required to display the pattern. To compute the phonetic identity of the pattern, time is needed, and this is provided by the segmentally redundant vowel. Vowels are the rests between consonants.

Rapid consonantal gestures cannot carry the melody and dynamics of the voice. The segmental and suprasegmental loads are therefore divided over consonant and vowel--the first, with its poor auditory store, taking the bulk of the segmental load, and the second taking the suprasegmental load. There emerges the syllable, a symbiosis of consonant and vowel, a structure shaped by the articulatory and auditory capacities of its user, fitted to, defining, and making possible linguistic and paralinguistic communication.

Cerebral Specialization for Syllable Perception

The distinctive acoustic structure of the syllable into which the speaker encodes consonant and vowel seems to call for a specialized neurophysiological decoding mechanism in the listener. Evidence for the operation of such a mechanism first came from the dichotic listening studies mentioned above (Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970). One question that these studies tried to answer was whether the mechanism was specialized for both auditory and phonetic analysis of the syllable or for phonetic analysis alone. I will not review the evidence here, but simply state our conclusion that "while the auditory system common to both hemispheres is equipped to extract the auditory parameters of a speech signal, the dominant hemisphere may be specialized for the extraction of linguistic features from these parameters" (Studdert-Kennedy and Shankweiler, 1970:594).

As we shall see shortly, recent evidence suggests that this conclusion may not be correct: the left hemisphere may be specialized for both auditory and phonetic analysis. First, however, we should note that Wood (1973a; also Wood, Goff, and Day, 1971) has provided impressive support for the conclusion in a study of the evoked potential correlates of phonetic perception. He synthesized two stop-vowel syllables, /ba/ and /ga/, which differed only in the extent and direction of their second and third formant transitions, the acoustic cues to their phonetic identities. He synthesized each at two fundamental frequencies: 104 Hz (low) and 140 Hz (high). From these syllables he constructed two types of test. In the first, fundamental frequency was held constant and the syllables were presented binaurally in random order: subjects identified each syllable phonetically, as fast as possible, by pressing one of two buttons. In the second type of test, phonetic identity was held constant, while fundamental frequency varied: subjects identified the fundamental frequency of each syllable as high or low. Both types of test therefore contained tokens of the same syllables, identified by pressing the same button with the same finger. During the tests, electrical activity was recorded from a central and a temporal scalp location over both left and right hemispheres. Evoked potentials were averaged and compared at each scalp location for the preresponse periods during presentation of identical syllables in the two tasks. Notice that the only possible source of variation in the EEG compared was in the task carried out by the subjects while the records were taken.

Statistical tests revealed significant differences between left-hemisphere records for the phonetic and fundamental frequency tasks at both locations. No significant differences appeared for either of the right-hemisphere locations. Furthermore, when the "speech" task called for identification of the isolated formant transitions of the two syllables--acoustic patterns which carry all the information necessary for phonetic identification, but which, lacking a following vowel, are not heard as speech--there were no significant left-hemisphere differences between records for "speech" and nonspeech tasks. The previously observed differences cannot therefore have been due to auditory analysis of the information-bearing formant transitions, but must presumably be attributed to phonetic interpretation of the auditory patterns. The experiments leave little doubt that different neural events occur in the left hemisphere, but not in the right hemisphere, during phonetic, as opposed to auditory, analysis of the same acoustic signal.

In other words, the language hemisphere does indeed appear to be specialized for phonetic interpretation (and, presumably, higher language functions), but not for auditory analysis of speech. This might seem to imply that the physical vehicle of the phonetic message is a matter of indifference. Superficial support for this view comes from two further sources. First, Papçun, Krashen, Terbeek, Remington, and Harshman (1974; also Krashen, 1972), have shown that experienced Morse code operators, identifying both individual letters and words presented dichotically, show a significant right-ear advantage. Second, Kimura (see Kimura and Durnford, in press) and others have repeatedly shown a right-field (left-hemisphere) advantage for tachistoscopically presented letters. If both Morse code and printed letters can invoke left-hemisphere processing, there might seem to be little reason to claim any special status for speech.

Nonetheless, there are solid grounds for making this claim. First, several studies have suggested that the left hemisphere is specialized for extracting acoustic features important in speech. Halperin, Nachshon, and Carmon (1973) have shown that the dichotic ear advantage shifts from left to right as the number of transitions in brief, temporally patterned sound sequences increases. Among their stimuli were permutations of three long (400 msec) and short (200 msec) sound bursts similar to the patterns used in Morse code. Their results therefore fit neatly with those of Krashen (1972) who found that naive subjects have a right-ear advantage for dot-dash sequences no more than seven units long. Taken together, the two studies suggest left-hemisphere specialization for judging duration and temporal pattern. Both studies have the weakness that subjects were asked to label the sequences, a process that might well invoke left-hemisphere control.

This objection is not decisive since arbitrary labeling of isolated formant transitions in Wood's (1973) study did not evoke the left-hemisphere potentials of phonetic labeling. Nonetheless, the weakness is avoided in some recent experiments by Cutting (in press). In one of these he constructed two-formant patterns identical with patterns signaling /bV/ or /dV/ except that their first-formant transitions fell rather than rose along the frequency scale, producing a phonetically impossible sound that subjects did not recognize as speech. In a nonlabeling dichotic recognition task with these stimuli, subjects gave a right-ear advantage of the same magnitude as for the normal CV syllables also used in the study. Cutting concludes from this and other experiments that the left hemisphere may be specialized for auditory analysis of speech.

But why then did the isolated transitions of Wood (1973) yield no left-hemisphere effect? The answer to this may be that the speech auditory analyzer is engaged not simply by acoustic features, but by features distributed over a signal of some minimum duration (such as that of a stressed syllable). Here the work of Wollberg and Newman (1972) on squirrel monkey is suggestive. They made single-cell recordings from cortical neurons responsive to the species' "isolation peep." A normal pattern of neuronal response occurred only if the entire "peep" was presented. Perhaps it is not far-fetched to suppose that the human cortex is supplied with sets of acoustic detectors tuned to speech, each inhibited from output to the phonetic system in the absence of collateral response in other detectors.

Be that as it may, the evidence for specialized left-hemisphere auditory analysis is, at best, preliminary and, in any case, not essential to the claim of special status for speech. Nor, indeed, is any form of speech-specific auditory analysis, whether unilateral or bilateral. Certainly, the accumulating evidence for specialized acoustic property detectors (Cutting and Eimas, this conference) is important and may even be decisive. But the initial strength of the claim comes from the distinctive structure of the syllable. The underlying phonological elements that determine this structure are common and peculiar to all languages. And recovery of those elements, whether from alphabet, optophonic light pattern, Morse code, or the neural display of an auditory system, engages mechanisms in the language hemisphere. The syllable is the structure on which the hierarchy of language is raised.

REFERENCES

Barclay, J. R. (1972) Noncategorical perception of a voiced stop. Percept. Psychophys. 11, 269-274.

Crowder, R. G. (1971) The sound of vowels and consonants in immediate memory. J. Verbal Learn. Verbal Behav. 10, 587-596.

Crowder, R. G. (1972) Visual and auditory memory. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Crowder, R. G. (1973) Precategorical acoustic storage for vowels of short and long duration. Percept. Psychophys. 13, 502-506.

Crowder, R. G. and J. Morton. (1969) Precategorical acoustic storage (PAS). Percept. Psychophys. 5, 365-373.

Cutting, J. E. (in press) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys.

Cutting, J. E. and P. Eimas. (this conference) Phonetic feature analyzers and the processing of speech in infants. [Also in Haskins Laboratories Status Report on Speech Research SR-37/38 (this issue).]

Darwin, C. J. and A. D. Baddeley. (in press) Acoustic memory and the perception of speech. Cog. Psychol.

Dorman, M. (1973) Discrimination of intensity differences on formant transitions in and out of syllable context. Haskins Laboratories Status Report on Speech Research SR-33, 13-18.

Fromkin, Victoria A. (1971) The nonanomalous nature of anomalous utterances. Language 47, 27-52.

Fry, D. B., A. S. Abramson, P. D. Eimas, and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Lang. Speech 5, 171-189.

Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute (University of Tokyo) 28, 67-73.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute (University of Tokyo) 29, 207-214.

Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift of ear superiority in dichotic listening to temporally patterned verbal stimuli. J. Acoust. Soc. Amer. 53, 46-50.

Kimura, D. (1961a) Some effects of temporal lobe damage on auditory perception. Canad. J. Psychol. 15, 156-165.

Kimura, D. (1961b) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.

Kimura, D. and M. Durnford. (in press) Normal studies on the function of the right hemisphere in vision. In Hemisphere Function in the Human Brain, ed. by S. J. Dimond and J. G. Beaumont. [London: Paul Elek (Scientific Books) Ltd.].

Krashen, S. (1972) Language and the left hemisphere. Working papers in Phonetics (Phonetics Laboratory, University of California, Los Angeles) 24.

Ladefoged, P. (1967) Three Areas of Experimental Phonetics. (New York: Oxford University Press).

Lane, Harlan L. (1965) The motor theory of speech perception: A critical review. Psychol. Rev. 72, 275-309.

Liberman, A. M., K. S. Harris, Joanne Kinney, and H. Lane. (1961) The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. 61, 379-388.

Liberman, A. M., I. G. Mattingly, and M. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (New York: Wiley).

Malmberg, B. (1955) The phonetic basis for syllable division. Studia Linguistica 9, 80-87.

Mattingly, I. G. (this conference) The human aspects of speech. [Also in Haskins Laboratories Status Report on Speech Research SR-37/38 (this issue).]

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Milner, B., L. Taylor, and R. W. Sperry. (1968) Lateralized suppression of dichotically presented digits after commissural section in man. Science 161, 184-185.

Papçun, G., S. Krashen, D. Terbeek, R. Remington, and R. Harshman. (1974) Is the left hemisphere specialized for speech, language, and/or something else? J. Acoust. Soc. Amer. 55, 319-327.

Pike, Kenneth L. (1943) Phonetics. (Ann Arbor, Mich.: University of Michigan Press).

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. thesis, University of Michigan. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.

Pisoni, D. B. (in press) Auditory short-term memory and vowel perception. Mem. Cog.

Pisoni, D. B. and Joan H. Lazarus. (in press) Categorical and noncategorical modes of speech perception. Percept. Psychophys.

Pisoni, D. B. and J. Tash. (in press) Reaction times to comparisons within and across phonetic categories: Evidence for auditory and phonetic levels of processing. Cog. Psychol. [Also in Haskins Laboratories Status Report on Speech Research SR-34, 77-88 (1973).]

Sachs, R. M. (1969) Vowel identification and discrimination in isolation vs. word context. Quarterly Progress Report (Research Laboratory of Electronics, Massachusetts Institute of Technology) QPR-93, 220-229.

Sawusch, J. R. and D. B. Pisoni. (1973) Category boundaries for speech and nonspeech sounds. Paper presented at the 86th meeting of the Acoustical Society of America, Los Angeles, Calif., November.

Shankweiler, D. P. and M. Studdert-Kennedy. (1967) Identification of consonants presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Sparks, R. and N. Geschwind. (1968) Dichotic listening in man after section of neocortical commissures. Cortex 4, 3-16.

Spellacy, F. and Sheila Blumstein. (1970) The influence of language set on ear preference in phoneme recognition. Cortex 6, 430-439.

Stetson, R. H., (1951) Motor Phonetics. (Amsterdam: North-Holland).

Stevens, Kenneth N. (1968) On the relations between speech movements and speech perception. Z. Phon., Sprachwiss. Komm. 21, 102-106.

Studdert-Kennedy, M. (in press-a) The perception of speech. In Current Trends in Linguistics, ed. by T. A. Sebeok. (The Hague: Mouton).

Studdert-Kennedy, M. (in press-b) Information processing in phonetic perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (Springfield, Ill.: C. C Thomas).

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper. (1970) The motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Sweet, H. (1877) Handbook of Phonetics. (Oxford: Clarendon). (Also, College Park, Md.: McGrath, 1970.)

Weiss, M. S. and A. S. House. (1973) Perception of dichotically presented vowels. J. Acoust. Soc. Amer. 53, 51-58.

Wollberg, Z. and J. D. Newman. (1972) Auditory cortex of squirrel monkey: Response patterns of single cells to species specific vocalizations. Science 175, 212-214.

Wood, C. C. (1973a) Levels of processing in speech perception: Neurophysiological and information-processing analyses. Ph.D. dissertation, Yale University. (Also Supplement bound with Haskins Laboratories Status Report on Speech Research SR-35/36.)

Wood, C. C. (1973b) Parallel processing of auditory and phonetic information in speech perception. J. Acoust. Soc. Amer. 55, 435(A).

Wood, C. C., W. R. Goff, and R. S. Day. (1971) Auditory evoked potentials during speech perception. Science 173, 1248-1251.

The Evolution of Speech and Language*

Philip Lieberman[+]
Haskins Laboratories, New Haven, Conn.

Although the view that human language is "unique" and is disjoint from the communications systems of all other animals is still current (Lenneberg, 1967), the research of the past century has demonstrated that Charles Darwin's theory of evolution through natural selection is essentially correct. Human language can be no more disjoint from the communications systems of other living animals than human respiration or human locomotion. The apparent uniqueness of human language, like the apparent uniqueness of fully bipedal locomotion, merely reflects the fact that the intermediate forms are extinct.

Human locomotion and human language both can be viewed as the result of gradual processes that evolved from phylogenetically simpler hominid ancestors. A human characteristic like bipedal locomotion structures virtually all aspects of human behavior. Tool use and tool manufacture, for example, are possible in Homo sapiens because our hands are free. Tool use and tool manufacture, of course, crucially involve the presence of cognitive factors. Without the human brain, bipedal locomotion would not be that useful. The evolution of both bipedal locomotion and the human brain mutually reenforced the evolution of the behavioral patterns of tool use and tool manufacture, which, in turn, placed greater selective advantages on both bipedal locomotion and enhanced cognitive abilities. It thus is both necessary and meaningful to discuss the evolution of human characteristics like bipedal locomotion and language in terms of the different factors that may have structured the selective factors resulting in the retention of the mutations that ultimately created Homo sapiens. These factors also are, and have been, operant in the evolution of other species. We thus can form and test hypotheses concerning the nature of human language and speech using data derived from other species.

I will discuss some of the factors that may be involved in the evolution of human language. These factors are necessarily linked; the presence of one particular factor is not, in itself, an explanation of the evolution of language. The absence, or lack of development, of one factor or another for modern Homo

---

*Paper presented at a National Institute of Child Health and Human Development conference, "The Role of Speech in Language," held at Columbia, Md., October 1973; to be published in the conference proceedings, ed. by J. F. Kavanagh and J. E. Cutting (Cambridge, Mass.: MIT Press).

[+] Also University of Connecticut, Storrs. After 1 July 1974 at Brown University, Providence, R. I.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

sapiens would imply the presence of intermediate grades of language relative to the language of present-day humans. As Darwin (1859) pointed out, evolution proceeds in small steps. I will propose a model that involves the evolution of a number of interrelated factors that gradually derive hominid linguistic ability. I will necessarily have to limit the discussion of each factor, and the list of factors obviously will not be complete, but I will discuss some of the data that make each factor part of a scientific theory, a theory that can be tested and extended.

Factor 1. Speech and Language

Since the focus of this conference is the relationship between speech and language, I will start with these factors, though I do not intend to claim that language is impossible without speech. Human language appears to involve closely the constraints of human speech. However, as I will try to show in the discussion of some of the other factors, other forms of language are possible without the presence of the particular characteristics of human speech.

The special link between human speech and human language was recognized in the pioneering 19th century studies of Broca (1861) and Wernicke (1874). Broca found that lesions in a small area of the brain situated near the motor cortex in the left, dominant hemisphere of the brain impaired speech production and writing. The victims of the "aphasia" could still move their tongues, lips, etc. In some instances they could sing, but they had difficulty when they either spoke or wrote. Lesions in the area of the brain that has come to be known as Broca's area essentially interfere with the organization of the articulatory maneuvers that produce speech, the "programs," as well as the written symbols that represent speech. Wernicke in 1874 described and localized the complementary aspect of aphasia. He located an area of the brain near the auditory centers of the left, dominant hemisphere. Lesions in this area produced an aphasia in which the victim left out words, used the wrong syntax, or "lost" the proper phonetic spellings of words. The victims of lesions in Wernicke's area essentially lose part of the "dictionary" and the grammar that every human carries about in his (or her) head. Both of these areas of the brain can be regarded as evolved additions to parts of the brain that deal with the production of sound (for Broca's area) and the perception of sounds (for Wernicke's area). Lesions in Wernicke's area clearly involve much more than the mere perception of sound, just as lesions in Broca's area involve much more than the ability simply to move the tongue, lips, jaw, etc. The total linguistic ability of the victim is impaired. The siting of these areas near the parts of the brain that are directly concerned with auditory signals suggests that special neural mechanisms evolved matched to, and as a consequence of, vocal communication.

We can test this hypothesis with data derived from the study of other species. In recent years a number of electrophysiological and behavioral studies have demonstrated that various animals have auditory detectors that are "tuned" to signals of interest to the animal. Even "simple" animals like crickets appear to have neural units that code information about the rhythmic elements of their mating songs. The calling songs of male crickets consist of stereotyped rhythmic pulse intervals and females respond to conspecific males by their songs (Hoy and Paul, 1973).

Similar results have been obtained in squirrel monkey (Saimiri sciureus). Wollberg and Newman (1972) recorded the electrical activity of single cells in the auditory cortex of awake monkeys during the presentation of recorded monkey vocalizations and other acoustic signals. Eleven calls, representing the major classes of this species' vocal repertoire, were presented along with tone bursts, clicks, and a variety of acoustic signals designed to explore the total auditory range of these animals. Extracellular unit discharges were recorded from 213 neurons in the superior temporal gyrus of the monkeys. More than 80 percent of the neurons responded to the tape-recorded vocalizations. Some cells responded to many of the calls that had complex acoustic properties. Other cells, however, responded to only a few calls. One cell responded with a high probability only to one specific signal, the "isolation peep" call of the monkey.

The experimental techniques necessary in these electrophysiological studies demand great care and great patience. Microelectrodes that can isolate the electrical signal from a single neuron must be prepared and accurately positioned. Most importantly, the experimenters must present the animals with a set of acoustic signals that explores the range of sounds that the animal would encounter in its natural state. Demonstrating the presence of neural mechanisms matched to the constraints of the sound-producing systems of particular animals is therefore a difficult undertaking. The sound-producing possibilities and behavioral responses of most "higher" animals make comprehensive statements on the relationship between perception and production difficult. We can only explore part of the total system of signaling and behavior. "Simpler" animals, however, are useful in this respect since we can see the whole pattern of the animal's behavior.

The behavioral experiments of Capranica (1965) and the electrophysiological experiments of Frishkopf and Goldstein (1963), for example, demonstrate that the auditory system of the bullfrog (Rana catesbeiana) has single units that are matched to the formant frequencies of the species-specific mating call. Bullfrogs are members of the class of Amphibia. Frogs and toads compose the order of Anura. They are the simplest living animals that produce sound by means of a laryngeal source and a supralaryngeal vocal tract. The supralaryngeal vocal tract consists of a mouth, a pharynx, and a vocal sac that opens into the floor of the mouth in the male. Vocalizations are produced in the same manner as in primates. The vocal folds of the larynx open and close rapidly, emitting puffs of air into the supralaryngeal vocal tract, which acts as an acoustic filter. Frogs can make a number of different calls (Bogert, 1960). These calls include mating calls, release calls, territorial calls which serve as warnings to intruding frogs, rain calls, distress calls, and warning calls. The different calls have distinct acoustic properties.

The mating call of the bullfrog consists of a series of croaks varying in duration from 0.6 to 1.5 sec. The interval between each croak varies from 0.5 to 1.0 sec. The fundamental frequency of the bullfrog croak is about 100 Hz. The formant frequencies of the croak are about 200 and 1400 Hz. Capranica (1965) generated synthetic frog croaks by means of a POVO speech synthesizer (Stevens, Bastide, and Smith, 1955). This is a fixed speech synthesizer designed to produce human vowels. It serves equally well for the synthesis of bullfrog croaks. In a behavioral experiment Capranica showed that bullfrogs responded to synthesized croaks so long as the croaks had energy concentrations at either or both of these frequencies (200 and 1400 Hz). The presence of acoustic energy at other frequencies inhibited the bullfrogs' responses (joining in a croak chorus).

Frishkopf and Goldstein (1963) in their electrophysiologic study of the bullfrog's auditory system found two types of auditory units. They found cells in units in the eighth cranial nerve of the anesthetized bullfrog that had maximum sensitivity to frequencies between 1000 and 2000 Hz. They found other units that had maximum sensitivity to frequencies between 200 and 700 Hz. The units that responded to the lower frequency range, however, were inhibited by appropriate acoustic signals. Maximum response occurred when the two units responded to time-locked pulse trains, at rates of 50 and 100 pulses per sec, that had energy concentrations at or near the formant frequencies of bullfrog mating calls. Adding acoustic energy between the two formant frequencies at 500 Hz inhibited the responses of the low-frequency single units.

The electrophysiologic, behavioral, and acoustic data are complementary. Bullfrogs have auditory mechanisms structured to respond specifically to the bullfrog mating call. Bullfrogs don't simply respond to any sort of acoustic signal as though it were a mating call. They respond only to particular calls that can be made only by male bullfrogs, and they have neural mechanisms structured in terms of the species-specific constraints of the bullfrog sound-producing mechanism. Capranica tested his bullfrogs with the mating calls of 34 other species of frogs. The bullfrogs responded only to bullfrog calls; they ignored all other mating calls. The croaks must have energy concentrations equivalent to those produced by both formant frequencies of the bullfrogs' supralaryngeal vocal tract. The stimuli furthermore must have the appropriate fundamental frequency.

The bullfrog has one of the simplest forms of sound-making systems that can be characterized by the Source-Filter Theory of sound production (Fant, 1960; to be discussed more fully below). His perceptual apparatus is demonstrably structured in terms of the constraints of his sound-producing apparatus and of the acoustic parameters of the Source-Filter Theory, the fundamental frequency and formant frequencies.

Factor 2. Plasticity and the Evolution of Human Speech

Frogs are rather simple animals but they nonetheless have evolved different species-specific calls. Some of the 34 species whose mating calls failed to elicit responses from Rana catesbeiana were closely related. Others were more distantly related. Clearly, natural selection has produced changes in the mating calls of Anuran species. The neural mechanisms for the perception of frog calls are at the periphery of the auditory system. They apparently are not very plastic since Capranica was not able to modify the bullfrogs' responses over the course of an 18-month interval. Despite this lack of plasticity, frogs have evolved different calls in the course of their evolutionary development.

Primates appear to have more flexible and plastic neural mechanisms for the perception of their vocalizations. Recent electrophysiological data (Miller, Stutton, Pfingst, Ryan, and Beaton, 1972) show that primates like rhesus monkey (Macaca mulata) will develop neural detectors that identify signals important to the animal. Receptors in the auditory cortex responsive to a 200 Hz sine wave were discovered after the animals were trained by the classic methods of conditioning to respond behaviorally to this acoustic signal. These neural detectors could not be found in the auditory cortex of untrained animals. The auditory system of these primates thus appears to be plastic. Receptive neural devices can be formed to respond to acoustic signals that the animal finds useful.

Factor 3. Special Supralaryngeal Vocal Tract Anatomy

Modern man's speech-producing apparatus is quite different from the comparable systems of living nonhuman primates (Lieberman, 1968; Lieberman, Klatt, and Wilson, 1969; Lieberman, Crelin, and Klatt, 1972). Nonhuman primates have supralaryngeal vocal tracts in which the larynx exits directly into the oral cavity (Negus, 1949). In the adult human the larynx exits into the pharynx. The only function for which the adult human supralaryngeal vocal tract appears to be better adapted is speech production. Understanding the anatomical basis of human speech requires that we briefly review the Source-Filter Theory of speech production (Fant, 1960). Human speech is the result of a source, or sources, of acoustic energy being filtered by the supralaryngeal vocal tract. For voiced sounds, that is, sounds like the English vowels, the source of energy is the periodic sequence of puffs of air that pass through the larynx as the vocal cords (folds) rapidly open and shut. The rate at which the vocal cords open and close determines the fundamental frequency of phonation. Acoustic energy is present at the fundamental frequency and at higher harmonics. The fundamental frequency of phonation can vary from about 80 Hz for adult males to about 500 Hz for children and some adult females. Significant acoustic energy is present in the harmonics of fundamental frequency to at least 3000 Hz. The fundamental frequency of phonation is, within wide limits, under the control of the speaker who can produce controlled variations by changing either pulmonary air pressure or the tension of the laryngeal muscles (Lieberman, 1967). Linguistically significant information can be transmitted by means of these variations in fundamental frequency as, for example, in Chinese where these variations are used to differentiate among words.

The main source of phonetic differentiation in human language, however, arises from the dynamic properties of the supralaryngeal vocal tract acting as an acoustic filter. The length and shape of the supralaryngeal vocal tract determines the frequencies at which maximum energy will be transmitted from the laryngeal source to the air adjacent to the speaker's lips. These frequencies, at which maximum acoustic energy will be transmitted, are known as formant frequencies. A speaker can vary the formant frequencies by changing the length and shape of his supralaryngeal vocal tract. He can, for example, drastically alter the shape of the airway formed by the posterior margin of his tongue body in his pharynx. He can raise or lower the upper boundary of his tongue in his oral cavity. He can raise or lower his larynx and retract or extend his lips. He can open or close his nasal cavity to the rest of the supralaryngeal vocal tract by lowering or raising his velum. The speaker can, in short, continually vary the formant frequencies generated by his supralaryngeal vocal tract. The acoustic properties that, for example, differentiate the vowels [a] and [i] are determined solely by the shape and length differences the speaker's supralaryngeal vocal tract assumes in articulating these vowels. The situation is analogous to the musical properties of a pipe organ, where the length and type (open or closed end) of pipe determines the musical quality of each note. The damped resonances of the human supralaryngeal vocal tract are, in effect, the formant frequencies. The length and shape (more precisely the cross-sectional area as a function of distance from the laryngeal source) determine the formant frequencies.

The situation is similar for unvoiced sounds where the vocal cords do not open and close at a rapid rate, releasing quasiperiodic puffs of air. The source of acoustic energy in these instances is the turbulence generated by air rushing through a constriction in the vocal tract. The vocal tract still acts as an acoustic filter but the acoustic source may not be at the level of the larynx as, for example, in the sound [s] where the source is the turbulence generated near the speaker's teeth.

The anatomy of the adult human supralaryngeal vocal tract permits modern man to generate supralaryngeal vocal-tract configurations that involve abrupt discontinuities at its midpoint. These particular vocal-tract shapes produce vowels like [a], [i], and [u], which have unique acoustic properties.[1] The acoustic properties of these sounds minimize the problems of precise articulatory control. A speaker can produce about the same formant frequencies for an [i], for example, while he varies the position of the midpoint area function discontinuity by 1 or 2 cm (Stevens, 1972). They are also sounds that are maximally distinct acoustically. They, moreover, are sounds that a human listener can efficiently use to establish the size of the supralaryngeal vocal tract he is listening to. This last property relates to Factor 1, the specialized speech decoding that characterizes human language. The reconstructions of the supralaryngeal vocal tracts of various fossil hominids that Edmund S. Crelin has made (Lieberman and Crelin, 1971; Lieberman et al, 1972; Lieberman, in press) indicate that some extinct hominids lacked the anatomical basis for producing these sounds, while other hominids appear to have the requisite anatomical specializations for human speech.

Factor 4. Syntactic Encoding and Decoding

There are three interrelated aspects to the cognitive abilities that underlie language: syntactic encoding and decoding, automatization, and logical ability. Syntactic encoding and decoding obviously involves the presence of neural mechanisms. Although we don't know very much about the workings of the brain, we don't have to know how the brain works to know what it does. A transformational grammar (Chomsky, 1957, 1964, 1968) is, among other things, a formal description of the syntactic encoding that is a characteristic of human language. Encoding in a more general sense seems to be a characteristic of other forms of human behavior.

A grammar to a linguist is not a set of prescriptive rules for writing sentences. A grammar is instead a formal description of some aspect of linguistic behavior. As Chomsky (1957:11) puts it:

> Syntactic investigation of a given language has as its goal the construction of a grammar that can be viewed as a device of some sort for producing the sentences of the language under analysis. More generally, linguists have been concerned with the problem of determining the fundamental underlying properties of successful grammars.

The fundamental property of grammar that Chomsky revealed is its "transformational syntax." Chomsky demonstrated that language must be viewed as a two-level process. Underlying the sequence of words that constitutes a normal, grammatical sentence is a "deep phrase marker" (Chomsky, 1964), which is closer to the logical level of analysis necessary for the semantic interpretation of a sentence. The transformational syntax is the "device" that restructures the deep, underlying level of language that is suited for semantic analysis, into the actual sentence that a person writes or speaks. The aspect of transformational syntax we want to stress is its encoding property, which is formally similar to the process of speech encoding (Liberman, 1970).

---

[1] As well as consonants like [g] and [k] which involve the velar region of articulation.

In Figure 1 we have presented a diagram that is essentially similar to the "parsing" or "constituent analysis" (Bloomfield, 1933) of traditional grammarians. The symbol S stands for sentence, NP for noun phrase, VP for verb phrase, V for verb, N for noun, and T for article. The diagram shows the syntactic relationships of the words of the sentence The man is old. The words the man, for example, constitute a noun phrase, the words is old constitute a verb phrase, which in turn is made up of a verb plus a second noun phrase. The word old constitutes the second noun phrase (the article of the second noun phrase reduces to an implied article). Diagrams of this sort are quite traditional. The first noun phrase could be called the subject of the sentence, the second, the object or predicate, etc. Semantic relationships are often "explained" by means of diagrams of this sort. The "actor-object" relationship, for example, is apparent in the diagram of the sentence Joe hit the man. The actor is the noun preceding the verb, the object the noun following the verb. We have simplified these diagrams and many of the details that a grammarian might find essential have been eliminated, but the essential facts and "explanatory" power of these diagrams have been preserved. Parsing is a "device" that formally "explains" some aspects of semantics, i.e., it reduces semantic analysis to a mechanical procedure. The noun to the left of the verb is the actor, that to the right of the verb is the object,

FIGURE 1

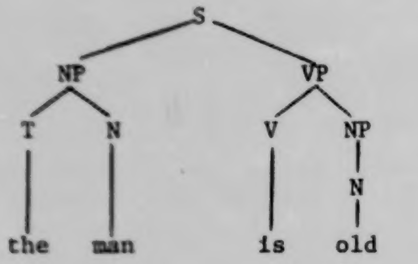
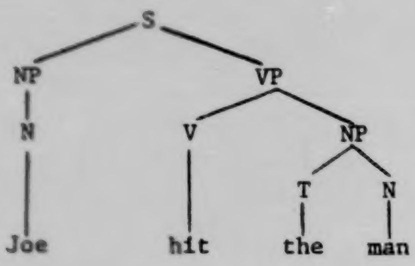


FIGURE 2



i.e., the noun acted on. The interesting thing about human language is that no one ever really utters sentences like *Joe hit the man* and *The man is old* when he wants to convey the information in the sentence *Joe hit the old man*. The two underlying deep phrase markers that would result in the simple sentences *The man is old* and *Joe hit the man* are encoded, i.e., scrambled together into one more complex sentence. The process is general and pervasive. The sentence *Joe hit the dirty old man who was wearing the red hat* would have underlying it a set of deep phrase markers that could have resulted in the sentences: *Joe hit the man*. *The man is old*. *The man is dirty*. *The man was wearing a hat*. *The hat is red*. It's much faster to utter the single complex sentence than the set of simpler sentences underlying it. The listener also doesn't have to keep track of the semantic referents and remember that we're talking about the same man in the first four simple sentences. All four repetitions of the word *man* are collapsed into a single *man* in the complex sentence. The two repetitions of the word *hat* are collapsed into a single *hat*. The complex sentence has fewer words and doesn't require keeping track of the semantic referents of the six "simple" sentences.

The transformational syntax can be regarded as the device that rearranges, deletes, and adds words to form the sentences of human language. The transformational syntax makes it impossible to sort out mechanically the semantic relationships of the words of complex sentences by using traditional sentence parsing. The "actor-acted on" relationship, for example, is semantically equivalent in the sentences *Joe hit Bill* and *Bill was hit by Joe*, though the words are on opposite sides of the verb. There are a number of reasons why traditional constituent grammars are not, in themselves, able to account for the properties of human languages (Chomsky, 1957, 1964; Postal, 1968), but it's enough to point out that they cannot account for the syntactic encoding that is characteristic of human language and for the complementary decoding that must take place when a listener or reader interprets a sentence.

Factor 5. Automatization

Human language involves rapidly executing complex sequences of articulatory movements or making equally complex perceptual decisions about the identity of particular sound segments. At a higher level, complex syntactic relationships must be determined. None of these processes is, however, what the speaker or listener is directly concerned with. The semantic content of the message is the primary concern of the speaker or listener. The sending and receiving processes are essentially automatic. No conscious thought is expended in the process of speech production, speech perception, or any of the syntactic stages that may intervene between the semantic content of the message and the acoustic signal. It is clear that "automatized" skills are not unique to human language. Other aspects of human activity, such as dance for example, involve similar phenomena. The novice dancer must learn the particular steps and movements characterizing a particular dance form. Once the steps have been learned they become automatized. The dance itself involves the complex sequences. Playing the violin, skiing, or driving a car all involve automatized behavior.

The bases for the automatized behavior that is a necessary condition for human language may reside in cross-modal transfers from other systems of hominid and primate behavior. Tool use, for example, requires a high degree of automatization. You can't stop to think how to use a hammer every time you drive a nail in. Hunting is perhaps a still stronger case. A successful hunter must be able to thrust a spear or throw a stone without pausing to think about the mechanics of spear thrusting or stone throwing. Natural selection would quickly favor the retention of superior automatization. Automatized behavior pervades all aspects of culture. Indeed a cultural response is, to a degree, a special case of automatized behavior. Electrophysiologic data derived from rhesus monkey demonstrates that automatization in primates involves establishing special pathways in the animal's motor cortex as the animal "learns" to perform a task (Evarts, 1973). Evarts observed the electrical activity of motor cortex neurons and the animal's muscles during the performance of learned hand movements. The animal's muscular activity when he learned to perform the task was extremely rapid. Its muscles acted within 30 to 40 msec, about twice as fast as it could have responded if it had to "think about" the task. Short response times like this usually are associated with reflex actions, but these short response times were the result of the animal's automatizing a response. The learned, automatized responses of simpler animals generally are not taken as tokens of the animal's "culture," but they nonetheless exist. The function of play in animals may indeed be to learn various patterns of automatized behavior germane to the animal's "culture." Puppies spend a lot of time staging mock battles, kittens stalk, etc. It wouldn't be difficult to devise appropriate experiments to explore the possible connection between play and automatization.

A special factor of automatized behavior may be that a "plastic" period is involved. It is comparatively easier to shape behavior during the plastic period. Afterwards, it is either impossible or more difficult for the animal or human to learn the automatized behavior. Puppies thus can be trained more readily than adult dogs. We're just beginning to appreciate some of the critical periods involved in learning various activities. Human newborns, for example, can be trained to walk alone about two months earlier than they normally do, if we take advantage of a critical period. Brief daily exercise of the walking reflexes that exist in human newborn leads to an earlier onset of walking alone (Zelazo, Zelazo, and Kolb, 1972). If a newborn infant is held under his arms and his bare feet are permitted to touch a flat surface, he will perform well-coordinated walking movements similar to those of an adult. This reflex normally disappears after about eight weeks. However, if the infant is actively exercised throughout this period, the reflex can be transferred intact from a reflex to a volitional action. Latent periods are quite important in the acquisition of human language (Lenneberg, 1967). All humans can readily learn different languages in their youth. They all appear to retain this ability to at least age 12 (Sachs, Lieberman, and Erickson, 1973). Most humans, however, can learn a foreign, i.e., unfamiliar, language only with great difficulty (or not at all) during adult life. There are, of course, exceptions to this rule and some adults are quite fortunate in retaining the ability to learn new languages with great facility. The same comments probably apply to learning to play the violin, tight-rope walking, etc., though no definitive studies have yet been made.

Factor 6. Cognitive Ability

Cognitive ability is a necessary factor in human language. Linguists often tend to assume that cognitive ability *is* linguistic ability. Indeed, since the time of Descartes the absence of human language in other animals has been cited as a "proof" of man's special status and of the lack of cognitive ability in all other species. Human language has been assumed to be a necessary condition for human thought. Conversely, the absence of human language has been assumed to be evidence of the lack of all cognitive ability.

The cognitive abilities traditionally associated with presumably "unique" human behavioral patterns like tool use and toolmaking have been observed in a number of different animals. Chimpanzees have often been observed using and making tools (Lawick-Goodall, 1972), but they are not the only primates who have been observed in the act of using and making tools. Beck (in press) reviews much of the evidence that shows tool use in other primates in their natural settings. Tool use has also been carefully documented in the sea otter (Kenyon, 1969). Sea otters float on their backs and use stones as anvils against which they break the shells of crustaceans. The sea otters will hold onto stones that are suitable anvils, tucking the stone under a flipper as they swim between meals. The sea otter thus not only uses a stone tool, but preserves it for future anticipated applications.

Tool use and toolmaking under less natural conditions has even been observed in birds. Laboratory-raised northern blue jays (Cyanocitta cristata) have been observed tearing pieces from pages of newspapers and using them as tools to rake in food pellets which were otherwise out of reach (Jones and Kamil, 1973). The toolmaking techniques that can be observed in living nonhuman animals are rather simple. The stone tools associated with the earliest known fossil hominids are, however, also rather simple. We'll discuss the cognitive implications of different toolmaking techniques, but it is clear that the tool-using and toolmaking behavior of many living animals is a reasonable approximation to the initial base on which natural selection acted in the gradual evolution of hominid behavior.

The linguistic ability of present-day chimpanzees also is evidence of the cognitive "base" that is present in living nonhuman animals. Chimpanzees do not have the speech-producing anatomy of modern Homo sapiens (Lieberman et al., 1972). They could not produce human speech even if they had the neural devices, localized in Broca's area, that organize the complex articulatory gestures of human speech. Chimpanzees, however, can be taught to use a modified version of American Sign Language. American Sign Language is not a method of "finger spelling" English words. It is instead a system that makes use of gestures that correspond to complete words, morphemes (e.g., past tense), or phrases (Stokoe, 1960). It has a different grammar than standard English and really is a different language with its own linguistic history. Chimpanzees taught this sign language communicate in a linguistic mode with human interlocutors (Gardner and Gardner, 1969; Fouts, 1973). They also can be observed communicating with other chimpanzees through sign language (Fouts, 1973). Other experimenters have taught chimpanzees to communicate with humans by means of plastic symbols (Premack, 1972) and by means of a computer keyboard (Rumbaugh, 1973). These experiments and observations demonstrate that chimpanzees can communicate in a linguistic mode. Chimpanzees, for example, are aware of what constitutes a "grammatical" syntactic construction (Rumbaugh, 1973). They conjoin words to form sentences such as I want apples and bananas, and they understand the principle of negation (Premack, 1972). They generalize the use of words, categorize in terms of semantic attributes, and use syntactic and logical constructs such as conditional sentences, Lucy read book if Roger tickle Lucy (Fouts, 1973). The chimpanzee's cognitive linguistic abilities are, at worst, restricted to some subset of the cognitive abilities available to humans. Chimpanzees may lack the syntactic encoding that must be formally described by a transformational syntax in human language. Definitive experiments investigating the syntax of chimpanzee communications using sign language have yet to be done, and we don't really know whether their sentences are syntactically encoded. The difference at the cognitive level may, however, be quantitative rather than qualitative.

It is important to note, at this point, that quantitative functional abilities can be the bases of behavioral patterns that are qualitatively different. I think that this fact is sometimes not appreciated in discussions of gradual versus abrupt change. A modern electronic desk calculator and a large general-purpose digital computer, for example, may be constructed using similar electronic logical devices and similar magnetic memories. The large general-purpose machine will, however, have 1,000 to 10,000,000 times as many logical and memory devices. The structural differences between the desk calculator and general-purpose machine may thus simply be quantitative rather than qualitative. The "behavioral" consequence of this quantitative difference can, however, be qualitative. The types of problems that one can solve on the general-purpose machine will differ in kind, as well as in size, from those suited to the desk calculator. The inherent cognitive abilities of humans and chimpanzees thus could be quantitative and still have qualitative behavioral consequences.

An Interactive Model for the Evolution of Human Language

I have discussed some of the factors that I think are relevant to the evolution of language and speech. The first hominid "languages" probably evolved from communication systems that resembled those of present-day apes. The social interactions of chimpanzees are marked by exchanges of facial and body gestures as well as vocalizations (Goodall, 1968). Chimpanzees also use tools, make tools, and engage in cooperative behavior (for example, hunting). All of these activities have been identified as factors that may have placed a selective advantage on the evolution of enhanced linguistic ability (Washburn, 1968; Hill, 1972).

Australopithecus africanus (Lieberman, 1973, in press) essentially has the same supralaryngeal vocal tract as present-day apes. This, however, still would allow *A. africanus* to establish a vocal language *if* other prerequisites were also present. *A. africanus* would have had to have had the motor skills and automatization necessary to produce the coordinated articulatory maneuvers that are necessary for the production of speech. Australopithecines were more advanced in relative brain size than any present-day ape, and, if external brain morphology means anything, they were more advanced in internal organization too. Quantities of shaped stones associated with early hominids have been recovered. [illegible]se stones probably were used, among other things, as projectiles (Leakey, 1971[illegible] The transference of patterns of "automatized" behavior from activities like toolmaking and hunting would have facilitated the acquisition of the motor skill necessary to make these sounds. Enhanced communicative ability would, in turn, facilitate the use of tools. The process would be circular, a positive feedback loop in which each step enhances the adaptive value of the next step. Particular neural capacities may initially not have been innately present. That is, they may not have been in place at birth like the auditory detectors of frogs, which don't appear to involve much, if any, learning. The plasticity of the Australopithecine auditory system, however, surely would have been at least as great as that of present-day rhesus monkeys, dogs, chaffinches, etc.

The initial language of the Australopithecines thus may have had a phonetic level that relied on both gestural and vocal components. The system may have become more elaborate as factors like tool use, toolmaking, and social interaction became more important. The ability to control rage and sex is one of the factors that makes human society possible (Hamburg, 1963). Language is probably one of the most important factors in reducing the level of agressive behavior in human society. Social control is as important a factor as hunting in the evolution of human society (Washburn, 1969). The level of interaction between mother and child which can be noted in the vocal and gestural communications of chimpanzee, in which the mother is the primary agency of socialization (Lawick-Goodall, 1972), is a good example of this source for the increased selective advantage of communication. As hominid evolution diversified and larger-brained hominids appeared in the Homo habilis/erectus lineage, the selective advantages of linguistic ability would have increased.

The final crucial stage in the evolution of human language would appear to be the development of the bent, two-tube supralaryngeal vocal tract of modern man. Figure 3 shows a divergence in the paths of evolution. Some hominids like the classic Neanderthal fossils appear to have retained the communication system that was typical of the Australopithecines, perhaps elaborating the system, but retaining a mixed phonetic level that relied on both gestural and vocal components (Lieberman and Crelin, 1971). Other hominids appear to have followed an evolutionary path resulting in almost complete dependence on the vocal component for language, relegating the gestural component to a secondary, "paralinguistic" function. The process would have been gradual, following from the prior existence of vocal signals in the linguistic communication of earlier hominids.

The bent supralaryngeal vocal tract that appears in forms like present-day Homo sapiens and the Es-Skhūl V fossil allows its possessors to generate acoustic signals that (1) have very distinct acoustic properties and (2) are easy to produce, being acoustically stable. These signals are in a sense optimal acoustic signals (Lieberman, 1970, 1973, in press). If vocal communications were already part of the linguistic system of early hominids, the mutations that extended either the range or efficiency of the signaling process would have been retained. At some later stage (that is, later with respect to the initial appearance of the bent, two-tube supralaryngeal vocal tract) the neural mechanisms necessary for the process of speech encoding would have evolved. The human-like supralaryngeal vocal tract would have initially been retained for the acoustically distinct and articulatorily stable signals that it could generate. The acoustic properties of the vowels [i] and [u] and the glides [y] and [w], which allow a listener to determine the size of a speaker's supralaryngeal vocal tract, would have preadapted the communication system for speech encoding.

The process of speech "decoding" appears to involve crucially the left hemisphere of the brain. When isolated vowels are, for example, presented dichotically to a human listener there is no right-ear advantage so long as the listener is responding to vowel stimuli that could have been produced by a single, unique vocal tract. If the vowel stimuli are instead derived from a set of different vocal tracts, a strong right-ear advantage is evident (Darwin, 1971). The listener has to make use of a perceptual recognition routine that normalizes the incoming signals in terms of the supralaryngeal vocal tracts that could have produced the particular stimuli. The neural modeling of this recognition routine apparently involves the left, dominant hemisphere of the listener's brain. The traditional mapping of areas like Broca's and Wernicke's areas in the left hemisphere of the brain reflects the result of a coherent evolutionary process in which the human brain evolved special, unique mechanisms structured in terms of the matched requirements of speech production and speech perception.

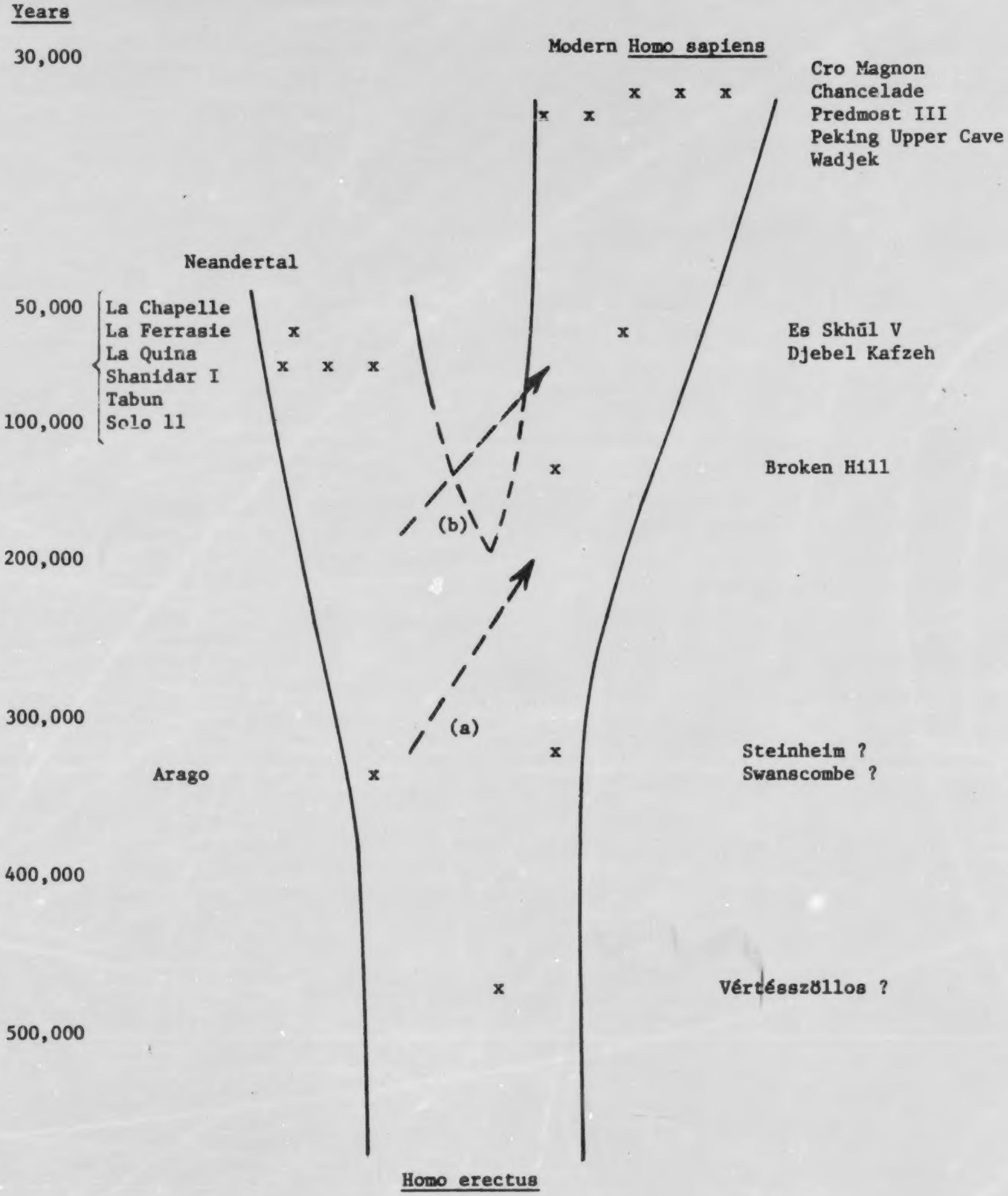

Years
30,000
50,000
100,000
200,000
300,000
400,000
500,000
Modern Homo sapiens
Cro Magnon
Chancelade
Predmost III
Peking Upper Cave
Wadjek
Neandertal
La Chapelle
La Ferrasie
La Quina
Shanidar I
Tabun
Solo 11
Es Skhūl V
Djebel Kafzeh
Broken Hill
(b)
(a)
Arago
Steinheim ?
Swanscombe ?
Vértésszöllos ?
Homo erectus

Figure 3: Tentative evolution of recent hominids with respect to human species.

The Uniqueness of Encoding

Although the speech of modern Homo sapiens is a fully encoded system, we can't assert dogmatically that other animals and, in particular, various fossil hominids, had completely unencoded systems of vocal communication. The acoustic basis of speech encoding rests in the fact that the pattern of formant frequency variation of the supralaryngeal vocal tract must inherently involve transitions. The shape of the supralaryngeal vocal tract cannot change instantaneously. If a speaker utters a syllable that starts with the consonant [b] and ends with the vowel [æ] his vocal tract must first produce the shape necessary for [b] and then gradually move towards the [æ] shape. Formant transitions thus have to occur in the [æ] segment that reflect the initial [b] configuration. The transitions would be quite different if the initial consonant were a [d]. The nonhuman supralaryngeal vocal tract can, in fact, produce consonants like [b] and [d]. Simple encoding could be established using only bilabial and dental consonant contrasts. The formant transitions would all be either rising in frequency in the case of [bæ] or falling in frequency for [dæ]. It probably would be quite difficult, if not impossible, to sort the various intermediate vowel contrasts that are possible with the nonhuman vocal tract, but a simple encoding system could be built up using rising and falling formant transitions imposed on a general, unspecified vowel [V]. The resulting language would have only one vowel [a claim that has often been made for the supposed ancestral language of Homo sapiens (Kuipers, 1960)]. The process of speech encoding and decoding and the elaboration of the vowel repertoire could build on vocal-tract normalization schemes that made use of sounds like [s], which also can provide a listener, or a digital computer program, with information about the size of the speaker's vocal tract. Vocal-tract normalizing information could also be derived perhaps by listening to a fairly long stretch of speech and then computing the average formant frequency range. The process would be slower than simply hearing a token of [i] or [u], but it would be possible.

There might have been a gradual path towards more and more encoding for all hominid populations as social structure and technology became more complex. The preadaptation of the bent, two-tube supralaryngeal vocal tract in some hominid populations would have, if this were true, provided an enormous selective advantage. In other words, there may not have been any single path towards the evolution of encoded speech. Fossil hominids like Neanderthal man may have had cognitive abilities equal to those of hominids like Es-Skhūl V. However, the absence of a preadapted, bent, two-tube vocal tract would have prevented them from generalizing the encoding principle.

Tool Use, Grammar, and Encoding

As we noted earlier, linguists often tend to view human language as though it were disjoint from all other aspects of human behavior. A linguistic grammar is essentially a formal description, or rather a formal abstraction, of certain aspects of language. Linguists, in general, would not think of applying the formal apparatus of a linguistic grammar to some other kind of human behavior. However, it is apparent that other aspects of human, and indeed of nonhuman, behavior can be described using the same formal apparatus. Reynolds, for example, who studied the play activity of young rhesus monkeys (Reynolds, 1972), found that rhesus monkeys have a number of stylized basic gestural patterns. These patterns are all quite short. They each consist of a particular body posture and facial expression. Some of the basic patterns involve movements and vocalizations. The basic play patterns are essentially "atomic" units that combine in certain regular ways to form play sequences. Some of the basic patterns occur only at the start of play sequences; they are "initiators." Others can occur only at the end of a play sequence; they are "terminators." Still other basic patterns occur within play sequences. The monkeys will break off play whenever a basic pattern occurs in the wrong position. There are, if we borrow the terminology of linguistics, "grammatical" play sequences and we can describe these sequences by means of "grammatical" rules.

The "grammatical" rules that appear to be appropriate for the description of rhesus monkey play sequences are those usually associated with constituent analysis, i.e., sentence parsing. Let's consider the following short example of sentence parsing drawn from Chomsky (1957). Consider the following set of grammatical "rules."

(1) S→NP + VP

(2) NP→T + N

(3) VP→V + NP

(4) T→the

(5a) N→man (5b) N→house (5c) N→ball

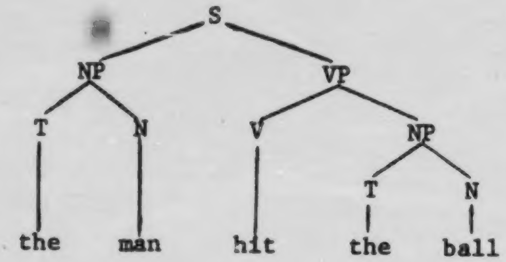
(6a) V→hit (6b) V→lost

Each rule X→Y is to be interpreted as the instruction "rewrite X as Y." We can call the sequence of operations that follows a "derivation" of the sentence, *The man hit the ball*, where the number at the right of each line in the derivation refers to the "rule" of the grammar used in constructing that line from the previous line.

S	
NP + VP	(1)
T + N + VP	(2)
T + N + V + NP	(3)
the + N + V + NP	(4)
the + man + V + NP	(5a)
the + man + hit + NP	(6a)
the + man + hit + T + N	(2)
the + man + hit + the + N	(4)
the + man + hit + the + ball	(5c)

Thus, the second line of the derivation is formed from the first line by rewriting sentence as NP + VP in accordance with rule (1), the third line is formed from the second line by using rule (2), etc. We could represent the derivation by means of the following diagram.

FIGURE 4



If we add a filter condition to the rules of the grammar it will mechanically derive a number of "grammatical" English sentences, e.g., *The man lost the house*, *The man hit the house*, etc. The filter condition states that no derivation shall be considered complete unless all of the alphabetic symbols are replaced by English words. The application of a particular rule in this grammar is contingent on only one fact--the left-hand symbol of a rule must be present on the last line of the derivation.

The grammar that we have discussed is what linguists call a "phrase structure" grammar. It's the formal embodiment of traditional sentence parsing. Phrase structure grammars in themselves cannot capture the encoded nature of the syntax of human language. Phrase structure rules, however, do have a role as a component of the grammar of human language (Chomsky, 1957, 1964). They have one formal property that, though it superfically appears trivial, is an important limitation of their explanatory power. A phrase structure rule can be applied in a derivation whenever the alphabetic symbol on the left of the rule appears on the last line of the derivation. *A phrase structure rule thus can apply to a line of a derivation without considering its past history*.

After digressing on the play activity of rhesus monkey and on phrase structure rules, we can now return to the question of the language of Neanderthal hominids. In fact, we have not really been digressing since the point that we want to make is that we can apply the "rules" of grammar to the analysis of some of the artifacts of Neanderthal culture, the stone tools and toolmaking techniques.

Stone Toolmaking Techniques and Encoding

The Paleolithic, or Old Stone Age, encompasses a period of perhaps almost three million years. There are important differences in the types of stone tools found in different parts of this era. The first tools, which are associated with the Australopithecines and *Homo habilis*, are either unshaped stones or stones that have a flake or two taken off them. The tools become progressively more complex and their manufacture ultimately involved taking many, many chips out of the piece of stone that the toolmaker started with. We might think of a process in which toolmakers continued to refine the process of tool fabrication, making the chips smaller and more refined as time went on. The basic technique, however, would be unchanged though new modifications would be introduced. The process would simply become more refined.

The technique involved in making these tools is conceptually similar to the process of whittling on a stick. You start by making an initial chip, then a second, a third, etc. In making a particular chip you have to keep only two things in mind: (1) the last chip that you made, and (2) the final form of the tool that you're trying to make. The process formally reduces to the phrase structure grammar with a filter condition that we just discussed. The filter condition is formally equivalent to stating that you know what sort of tool you're aiming for. The phrase structure grammar formally embodies the fact that you only need to know the last "line of the derivation," i.e., the state of the tool blank at the instant that you chip it. You don't need to have a memory of the operations involved in getting to that stage.

We would be wrong in thinking that all stone tools involved the same technology. About 600,000 years ago a radically different stoneworking technology started. The Levallois flake tools (Bordes, 1968) are the result of a multistage process. The toolmaker first prepares a core (Figure 5), a process which involves a number of steps itself to produce the basic shape. Once the core is ready, the toolmaker switches his technique. He chips out complete flakes, each of which may serve as a completed tool, with every blow of his hammer. The Levallois toolmaking technique cannot be reasonably described by means of a phrase structure grammar. A transformational grammar which formally incorporates a memory is necessary. There is no simple invariant "last chip" at which the toolmaker abruptly stops preparing the core and switches to flaking off the final products. The toolmaker rather has to keep in mind a particular functional attribute of the striking platform which involves the entire upper surface of the core (Bordes, 1968:27, 28). The formal "grammatical" description of the process must also reflect this degree of abstraction, which cannot be keyed to the appearance of a single "alphabetic" symbol that represents a particular chip of stone.

Phrase structure grammars cannot formally account for the syntax of human language (Chomsky, 1957, 1964); they also cannot serve as grammars of the Levalloisian tool technique that is one of the characteristics of the culture of Neanderthal man. Transformational grammars, as we noted, introduce the concept of encoding into syntax. Although we cannot positively conclude that the grammar of the syntax of Neanderthal language had a transformational component, their Levalloisian stone tools suggest a degree of cognitive development that formally calls for a transformational grammar. Many other aspects of the culture of modern human populations need transformational descriptions if we attempt to derive a formal description. Marriage customs, for example, involve constraints on the lineages of both bride and groom that include a memory component. Death rituals involving funeral goods also implicitly require some knowledge of the former life and habits of the corpse.

The most likely assessment of the encoding abilities of Neanderthal man thus would be that language was encoded, but not nearly as encoded as modern *Homo sapiens*'. The development of the Neanderthal supralaryngeal vocal tract was not suitable for fully encoded speech. The neural structures of the brain that play so crucial a role in the perception of encoded speech in the dominant, left hemisphere of the brain would therefore probably not have been as well developed in Neanderthal man. Language, however, would exist though it would not be the language of modern *Homo sapiens*. Language, like other human attributes, appears to be the result of a gradual evolutionary process, and intermediate stages and

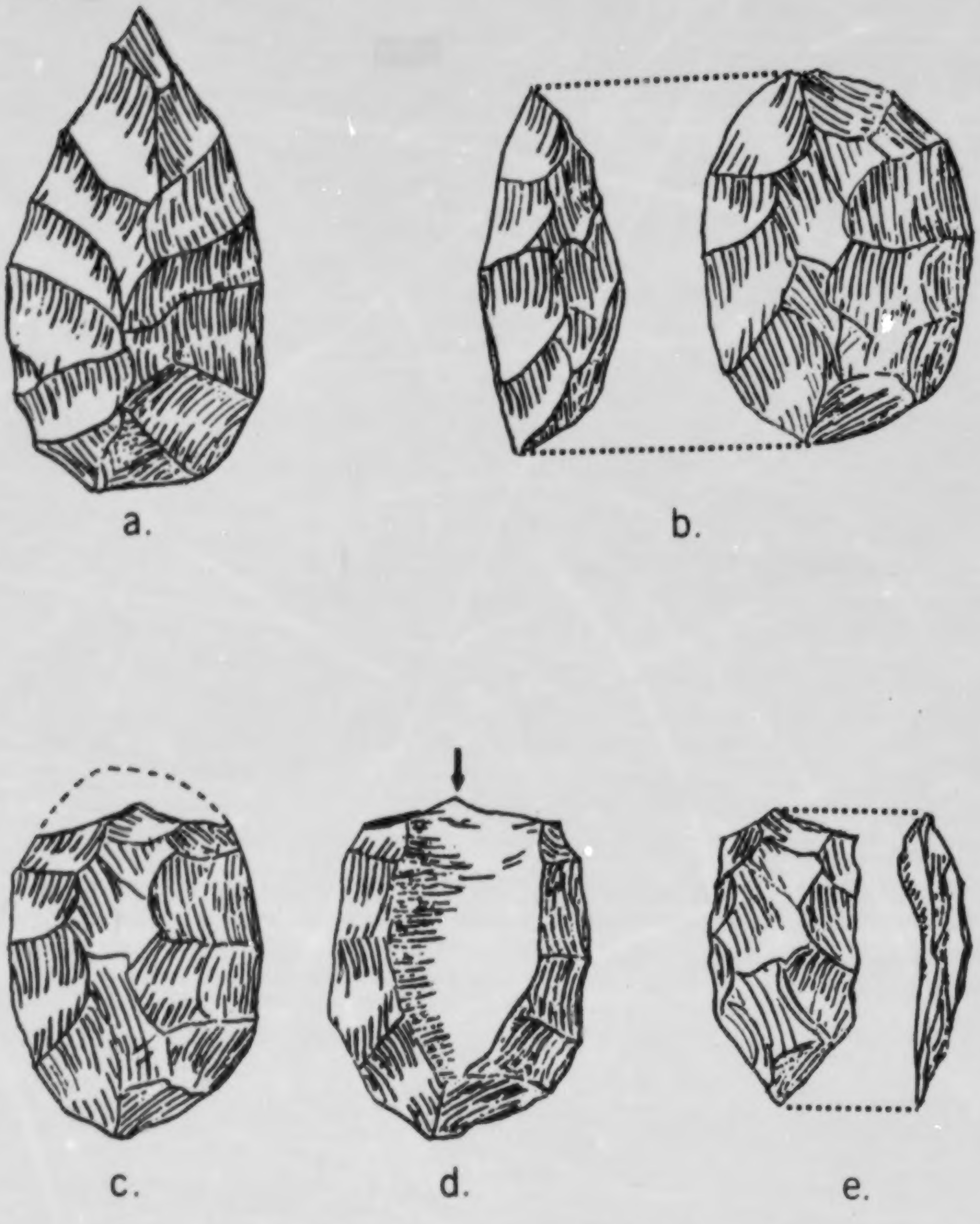


Figure 5: Paleolithic stone tools to illustrate an analogy between early toolmaking and syntax in language. (a) Handaxe or coup de poing. Lower Paleolithic of Africa, Europe, southwestern and southern Asia. Length of average specimens approximately 10 to 15 cm. Similarly chipped on opposite surface. (b)-(e) Stages in the manufacture of a Levallois flake from a prepared tortoise core. Late Lower Paleolithic and Middle Paleolithic, in many of same geographic areas as (a) above. Approximately similar scale. (e) The prepared tortoise core, back and side view; the flattish underside is also chipped. [After G. W. Hewes (this conference) "Some Comments on Mattingly's Paper and on Levallois Flake Tools."]

common underlying factors are to be expected in the languages and communications systems of extinct earlier hominids and of other living species.

REFERENCES

Beck, B. B. (in press) Primate tool behavior. In Proceedings of IXth International Congress of Anthropological and Ethnological Sciences, Chicago, Ill., 1973. (The Hague: Mouton).

Bloomfield, L. (1933) Language. (New York: Holt).

Bogert, C. M. (1960) The influence of sound on the behavior of amphibians and reptiles. In Animal Sounds and Communication, ed. by W. E. Lanyon and W. N. Tavolga. (Washington, D.C.: American Institute of Biological Sciences).

Bordes, F. (1968) The Old Stone Age. (New York: McGraw-Hill).

Broca, P. (1861) Nouvelle observation d'aphemie produite par une lesion de la moitie posterieure des deuxieme et troisieme ciconvolutions frontales. Bull. Soc. Anatom. Paris 6 (series 2), 398-407.

Capranica, R. R. (1965) The Evoked Vocal Response of the Bullfrog. (Cambridge, Mass.: MIT Press).

Chomsky, N. (1957) Syntactic Structures. (The Hague: Mouton).

Chomsky, N. (1964) Aspects of the Theory of Syntax. (Cambridge, Mass.: MIT Press).

Chomsky, N. (1968) Language and Mind. (New York: Harcourt).

Darwin, C. (1859) On the Origin of Species, facsimile edition. (New York: Atheneum).

Darwin, C. J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 386-392.

Evarts, E. V. (1973) Motor cortex reflexes associated with learned movement. Science 179, 501-503.

Fant, G. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Fouts, R. S. (1973) Acquisition and testing of gestural signs in four young chimpanzees. Science 180, 978-980.

Frishkopf, L. S. and M. H. Goldstein, Jr. (1963) Responses to acoustic stimuli from single units in the eighth nerve of the bullfrog. J. Acoust. Soc. Amer. 35, 1219-1228.

Gardner, R. A. and B. T. Gardner. (1969) Teaching sign language to a chimpanzee. Science 165, 664-672.

Goodall, J. (1968) A preliminary report on expressive movements and communication in the Gombe Stream chimpanzees. In Primates: Studies in Adaptation and Variability, ed. by P. Jay. (New York: Holt, Rinehart & Winston).

Hamburg, D. A. (1963) Emotions in the perspective of human evolution. In Expression of the Emotions in Man, ed. by P. Knapp. (New York: International Universities Press).

Hill, J. H. (1972) On the evolutionary foundations of language. Amer. Anthropol. 74, 308-317.

Hoy, R. R. and R. C. Paul. (1973) Genetic control of song specificity in crickets. Science 180, 82-83.

Jones, T. B. and A. C. Kamil. (1973) Toolmaking and tool-use in the northern blue jay. Science 180, 1076-1077.

Kenyon, K. W. (1969) The Sea Otter in the Eastern Pacific Ocean. (Washington, D.C.: U. S. Government Printing Office).

Kuipers, A. H. (1960) Phoneme and Morpheme in Kabardian. (The Hague: Mouton).

Lawick-Goodall, J. (1972) In the Shadow of Man. (New York: Macmillan).

Leakey, M. D. (1971) Olduvai Gorge, Vol. III. (Cambridge: Cambridge University Press).

Lenneberg, E. H. (1967) Biological Foundations of Language. (New York: Wiley).

Liberman, A. M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.

Lieberman, P. (1967) Intonation, Perception, and Language. (Cambridge, Mass.: MIT Press).

Lieberman, P. (1968) Primate vocalizations and human linguistic ability. J. Acoust. Soc. Amer. 44, 1574-1584.

Lieberman, P. (1970) Towards a unified phonetic theory. Ling. Inq. 1, 307-322.

Lieberman, P. (1973) On the evolution of human language: A unified view. Cognition 2, 59-94.

Lieberman, P. (in press) On the Origins of Language: An Introduction to the Evolution of Human Speech. (New York: Macmillan).

Lieberman, P. and E. S. Crelin (1971) On the speech of Neanderthal man. Ling. Inq. 2, 203-222.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. 74, 287-307.

Lieberman, P., D. H. Klatt, and W. A. Wilson. (1969) Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. Science 164, 1185-1187.

Miller, J. M., D. Sutton, B. Pfingst, A. Ryan, and R. Beaton. (1972) Single cell activity in the auditory cortex of rhesus monkeys: Behavioral dependency. Science 177, 449-451.

Negus, V. E. (1949) The Comparative Anatomy and Physiology of the Larynx. (New York: Hafner).

Postal, P. M. (1968) Aspects of Phonological Theory. (New York: Harper & Row).

Premack, D. (1972) Language in chimpanzee? Science 172, 808-822.

Reynolds, P. C. (1972) Play, language, and human evolution. Paper presented at the 1972 meeting of the American Association for the Advancement of Science, Washington, D.C.

Rumbaugh, D. M. (1973) Reading and sentence completion by a chimpanzee. Science 182, 731-733.

Sachs, J., P. Lieberman, and D. Erickson. (1973) Anatomical and cultural determinants of male and female speech. In Language Attitudes: Current Trends and Prospects, ed. by R. Shuy and R. Fasold. (Washington, D.C.: Georgetown University Press).

Stevens, K. N. (1972) Quantal nature of speech. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw-Hill).

Stevens, K. N., R. P. Bastide, and C. P. Smith. (1955) Electrical synthesizer of continuous speech. J. Acoust. Soc. Amer. 27, 207.

Stokoe, W. C., Jr. (1960) Sign Language Structure: An Outline of the Visual Communication System of the Deaf, Studies in Linguistics Occasional Paper 8. (Buffalo, N. Y.: University of Buffalo).

Washburn, S. L. (1968) The Study of Human Evolution. (Eugene, Ore.: Oregon State System of Higher Education).

Washburn, S. L. (1969) The evolution of human behavior. In The Uniqueness of Man, ed. by J. D. Roslansky. (Amsterdam: North-Holland).

Wernicke, C. (1874) Der Aphasische Symptomen-Complex. (Breslau: Franck & Weigert).

Wollberg, Z. and J. D. Newman. (1972) Auditory cortex of squirrel monkey: Response patterns of single cells to species-specific vocalizations. Science 175, 212-214.

Zelazo, P. R., N. A. Zelazo, and S. Kolb. (1972) "Walking" in the newborn. Science 176, 314-315.

Phonetic Feature Analyzers and the Processing of Speech in Infants*

James E. Cutting[+] and Peter D. Eimas[++]

Recently, Kaplan and Kaplan (1971) asked the question: "Is there such a thing as a prelinguistic child?" The traditional answer was a rather emphatic yes: for example, the first months of a child's life are generally characterized by nonlinguistic vocalizations (Jakobson, 1968; Lieberman, Crelin, and Klatt, 1972). Recent work, however, indicates that the answer to this question might be no; in fact, must be no.

Certainly one must be impressed with the child's rapid mastery of complex speech utterances and the rules for generating them (see Bloom, 1970; McNeill, 1970; Slobin, 1971; Menyuk, 1971; Brown, 1973). This research, however, has dealt with the speech production of children between the ages of 18 months and 5 years. Before that time, and certainly before the age of 12 months, the speech productions of a child are rather infrequent and erratic; semirandom babbling is the rule. Indeed, when the infant is very young there is evidence that his vocal tract is not even equipped to emit a repertoire of speech sounds that correspond to those found in the speech of normal adults (Lieberman, Harris, Wolff, and Russell, 1972). Since the ontogeny of language production in the infant appears to be severely constrained by the ontogeny of the vocal apparatus, the answer to the Kaplans' question may lie in the realm of speech perception rather than speech production.

Unlike other sensory processing systems of the neonate, his auditory capabilities are well developed. Wolff (1966) has shown that at two weeks the infant can tell the difference between a voice and other auditory sounds; Wertheimer (1961) has shown that neonates can localize sounds at birth; and there is evidence that the fetus responds to auditory stimulation several weeks before birth (Bernard and Sontag, 1947), perhaps even to the speech of the mother.[1] Since the young infant possesses a well-tuned auditory apparatus, it seems reasonable to devise an experimental situation in which the infant is asked if he

---

*Paper presented at a National Institute of Child Health and Human Development conference, "The Role of Speech in Language," held at Columbia, Md., October 1973; to be published in the conference proceedings, ed. by J. F. Kavanagh and J. E. Cutting (Cambridge, Mass.: MIT Press).

[+] Haskins Laboratories and Yale University, New Haven, Conn.

[++] Brown University, Providence, R. I.

[1] J. Bosma and D. Baker, personal communication.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

can perceive language events in a linguistic fashion. Two problems immediately arise: (a) what question should we ask, and (b) how should we ask it?

The appropriate linguistic question. It would be somewhat ludicrous to ask the infant to disambiguate a syntactically ambiguous sentence. Instead, it would be more appropriate to ask him about an earlier aspect of language processing (see Studdert-Kennedy, in press). An infant, for example, might be able to make a discrimination at the phonological level; or, perhaps more likely, he might be able to perceive the difference between a pair of phonemes which differ along a single phonetic feature. Since this appears to be a reasonable level of language to ask the infant about, it is necessary to consider which phonetic features the infant might be able to perceive.

One prominent feature is voicing. This feature separates the consonant phonemes in the following nonsense syllables: [ba] (as in bottle) from [pa], [da] from [ta], and [ga] from [ka]. A second candidate is place of articulation, a feature which distinguishes [ba] from [da] from [ga], and [pa] from [ta] from [ka]. Nasalization is a third phonetic feature which, for example, distinguishes [ma] from [ba]. Frication is a fourth likely feature, a dimension roughly separating [sa] from [ta].

In our first study (Eimas, Siqueland, Jusczyk, and Vigorito, 1971), we selected the feature of voicing. Voicing is a very stable phonetic feature for particular individuals within a given culture (Lisker and Abramson, 1964, 1967); it is universal in all languages, or nearly so (Lisker and Abramson, 1964, 1970; Abramson and Lisker, 1965, 1970); and it is quite prominent in the acoustic stream (Jakobson, Fant, and Halle, 1951). Also, voiced and voiceless consonants can be synthesized rather easily by a computer-driven parallel resonance synthesizer, such as that available at Haskins Laboratories (Mattingly, 1968). Moreover, they can be synthesized along a continuum with equal increments of acoustic change between the members of the stimulus array.

Since more is known about initial consonants than about final or medial consonants, the stimuli used in the first experiment and all subsequent ones were initial consonants in consonant-vowel (CV) nonsense syllables. The voiced-consonant syllables were all perceived as [ba] by adult listeners, and the voiceless-consonant syllables were perceived as [pa]. In natural speech, syllable-initial [b] and [p] phonemes are distinguished by the timing relationship between the release of the constriction at the lips, and the onset of the pulsing of the vocal folds. For an American English [ba] these two events happen very nearly at the same time. For [pa], however, there is a slight lag in the onset of the action of the vocal folds. Thus, the release of the constriction may occur 40 to 60 msec before the onset of voicing at the glottis.

Our linguistic question must be redefined as two questions. First, can the infant tell the difference between members of a cross-phoneme-boundary pair of stimuli; for example, [ba] from [pa]? If the answer is yes, a more sophisticated second question must then be asked: like the adult, can the infant not tell the difference between members of a within-phoneme-boundary pair; that is, two tokens of [ba] that have different voice onset times, or correspondingly, two different tokens of [pa]? For adults this peculiar capability is called categorical perception (see, among others, Liberman, 1957; Liberman, Harris, Kinney, and Lane, 1961; Abramson and Lisker, 1965; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Mattingly, Liberman, Syrdal, and Halwes, 1971; Pisoni, 1971, 1973).

Unlike most acoustic patterns, certain members of some speech continua appear to be discriminable only to the extent that the member stimuli can be labeled differently. If infants yield experimental results that are functionally parallel to those of adults, the inference is nearly irresistible that infants perceive speech.

The linguistic question appropriately posed. Knowing what to ask the infant is only half the battle; how to ask the question is the other and perhaps more difficult half. Here it is necessary to consider what the infant does, and what is important in his ecology.

Young infants sleep an inconvenient amount of time, and to find an awake and alert infant is in itself not a trivial problem. In order to measure his receptivity to events in his environment it is necessary to use one of the many indicants of the orienting response, some of which are found even in neonates (Kessen, Haith, and Salapatek, 1970). Successful measures of the infant's awareness of the environment and of environmental changes include visual fixation (e.g., Karmel, 1969), changes in heart rate (e.g., Graham and Jackson, 1970), and systematic changes in the EEG pattern (e.g., Molfese, 1972; Dorman, in press).

Yet another response system, which has high ecological validity and which plays a dominant role in the infant's earliest encounters with his environment, is the sucking response. Moreover, unlike heart rate and EEG recordings, sucking responses are highly visible and easily measured. Siqueland and DeLucia (1969) have used this response to great advantage in assessing the visual perception of infants. We have also used the sucking response, but our interest has been to evaluate the infant's perception of acoustic events.

In this experimental situation the infant is given a hand-held nipple upon which to suck. Instead of the nipple transducing nutrients it transduces pressure, which is in turn transformed into polygraphic and digital records of the sucking responses. Contingent on the sucking response is the presentation of an auditory stimulus, one of the members of the speech continua synthesized for the study.[2] Two different methodological criteria have been used for stimulus presentations. In one method, used by Eimas et al. (1971), the intensity of the stimulus is contingent on the rate of the infant response. While the rate of presentation is held constant, stimulus intensity is increased for rapid responding from an inaudible level to as much as 75 db sound pressure level against 63 db background noise. If the infant is not sucking at a high rate, the amplitude of the stimulus is systematically decreased.

In the second method, which is currently being used, the rate of the stimulus presentation is contingent on the rate of sucking. Stimulus presentation and sucking response are nevertheless not always related one-to-one. Each stimulus is 500 msec in duration, followed by a compulsory period of silence which is also 500 msec in duration. Thus, there is an irreducible refractory period of one second; though the infant may respond at a rate faster than one per second, the items are not presented at a rate greater than that. If, however, the infant sucks at a rate less than once per second, stimulus presentation rate exactly corresponds to response rate.

---

[2] In actuality, only high-amplitude sucking responses resulted in auditory stimulation. The amplitude was set for each infant individually, such that it yielded a baseline sucking rate between 20 and 30 responses per minute.

The infant quickly learns the relationship between the presentation of the stimulus and his sucking response, and he is quite willing to make from 200 to 600 sucking responses to listen to a particular stimulus during the course of a 10-minute experimental session. This remarkable effort to obtain stimulation is an impressive testament to the curiosity that we are born with.

Stimuli, procedure, and results. Shown at the top of Figure 1 is a sound spectrogram of an extremely prevoiced [ba], a stimulus in which the vibration of the vocal folds began 150 msec before the lip release was initiated. Plotting time against frequency, we see that the signal is easily divisible into three temporal segments: an initial low-frequency, low-amplitude steady-state voice bar which immediately precedes the release of the constriction; a 40-msec segment of formant transitions which increase in frequency; and a considerably longer segment of three steady-state formants which correspond to the vowel [a]. To vary stimuli along the dimension of voice onset time (VOT), several acoustic parameters must be changed. When a stimulus is varied between -150 msec VOT and 0 msec VOT, only the voice bar is altered. The duration of the voice bar denotes the appropriate negative value on the VOT continuum.[3] When voice onset follows the release the acoustic manifestation of VOT is changed. The onset of the first formant is cut back by the amount of difference between release and voice onset. For example, if a stimulus has a +40 msec VOT, the onset of the first formant (F1) is precisely at the beginning of the steady-state resonance; that is, there is no F1 transition. Note that the lower spectrogram in Figure 1, the stimulus [pa], has no F1 transition.

Acoustic changes are also revealed in the upper formants for all positive value voice onsets. F2 and F3 retain their shapes and frequencies, but they are excited by a different sound source. Before voice onset an aperiodic, hissing sound is created by a local turbulence near the point of constriction in the mouth (in this case the lips). After voice onset the upper formants attain their more accustomed appearance, driven by the periodic glottal source. Figure 1 displays spectrograms of three stimuli from this type of continuum whose VOT values are -150, +10, and +100 msec. Also shown are the overall amplitude envelopes for each stimulus.

The stimuli used in the first study in this series were those from a [ba]-[pa] continuum synthesized by Lisker and Abramson at the Haskins Laboratories. The VOT values were -20, 0, +20, +40, +60, and +80 msec. Since the [b]-[p] phoneme boundary in English is at about +25 msec VOT (Lisker and Abramson, 1970; Abramson and Lisker, 1973), the pair of stimuli whose values are +20 and +40 lie within different phoneme categories; the +20 msec stimulus is typically identified as [ba], and the +40 msec stimulus is identified as [da]. This pair is called the D pair since the members belong to different categories. Two other pairs of stimuli were S pairs since they belong within the same phoneme category. These pairs were -20 and 0, and +60 and +80. The members of the first pair are both identified by adults as [ba], and those of the second pair as [pa].

---

[3] Convention has it that when voicing onset precedes the release the interval is measured in negative VOT values, and when voicing onset follows the release the interval is measured in positive values (Abramson and Lisker, 1965).

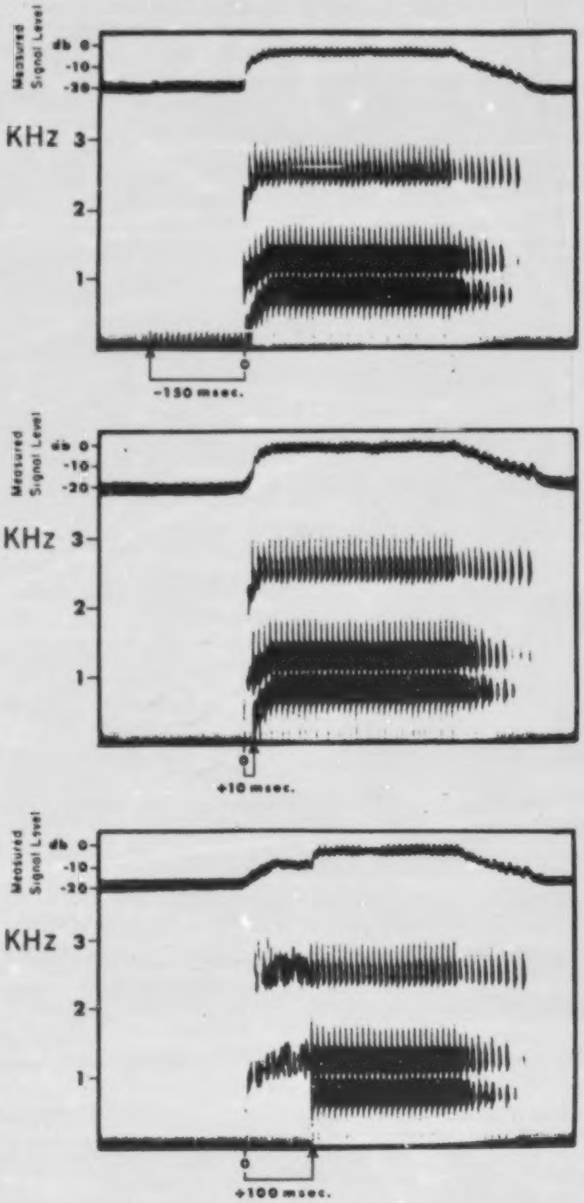

Three Conditions of Voice Onset Time
Synthetic Labial Stops

Figure 1: Three stimuli used in voice-onset time discrimination studies. In American English the first two are [ba], and the third is [pa]. (Adapted with permission from Abramson and Lisker, 1973.)

Before the experimental session begins, a baseline sucking response rate must be obtained from each infant against which all subsequent response rates can be measured. Afterwards the experiment begins its preshift stage, and stimuli are presented contingent on the infant's sucking responses. Within a few minutes the infant learns the association between stimulus and response, and increases his response rate to approximately 50 to 60 responses per minute. After the response peak occurs the infant's responses typically decrease in a dramatic fashion, the hallmark of adaptation. At least two minutes after the peak rate and a decrease of at least 20 percent, the postshift experimental phase begins and the infants' response rates diverge according to the stimuli that are presented.

Members of Group C, the control group, continue to listen and respond to the same stimulus that they heard in the previous portion of the experiment. Regardless of which VOT stimulus the infants listen to, their response rate continues to decrease and approach an asymptote. Unlike those in the control group, Group S and D infants experience a shift in stimulation. Those in Group S listen and respond to a different stimulus, but nonetheless a stimulus whose initial consonant is from the same phoneme category. Their response rate, like that of the control group, typically continues to decrease towards zero.

Infants in Group D, on the other hand, listen and respond to a new stimulus whose initial consonant belongs to a different phoneme category. The response rate of this group is quite different from that of the other groups. Instead of continuing to decline, their response rate increases markedly. These infants often maintain a higher-than-preshift response rate throughout the four-minute postshift period. A schematic representation of the response rates typical of the three groups is shown in Figure 2. In the initial study (Eimas et al., 1971) there was no difference between the postshift response functions for Group S and Group C. Group D, however, showed a significantly higher response rate than either of the other two groups. Furthermore, one-month and four-month infants in this group yielded essentially identical response functions.

The implication of these results is compelling: since the infant and the adult perceive some speech events in a similar manner, and since the infant has had only a very limited exposure to language, the mechanisms by which he perceives speech must be innate. Perhaps they are phonetic feature analyzers.

Some problems and their resolutions. Before amplifying this conclusion, however, it is necessary to consider a few problems. First, there are a few procedural problems. For example, in this type of experimental situation only about 40 to 50 percent of the infants make it through the entire session. Others cry, fall asleep, or their response rate decreases so rapidly before the experimental shift that it has fallen considerably below baseline, a situation in which their data must be ignored. This difficulty, although trying for the experimenter, is minor. There is no reason to believe that infants would differ in the manner that they perceive these stimuli; that is, there is no evidence that infants who do not fulfill the requirements for remaining in the experiment possess analyzing systems of a markedly different nature. They are just fussier.

Another problem arises when considering the age of the infants that can be used as subjects. The first study used one- and four-month-old infants for a very good reason. At younger than one month the infant will usually fall asleep before the 10-to-15 minute experimental session is complete. After four months the infant becomes too active, and begins to crawl out of the experimental

FIGURE 2

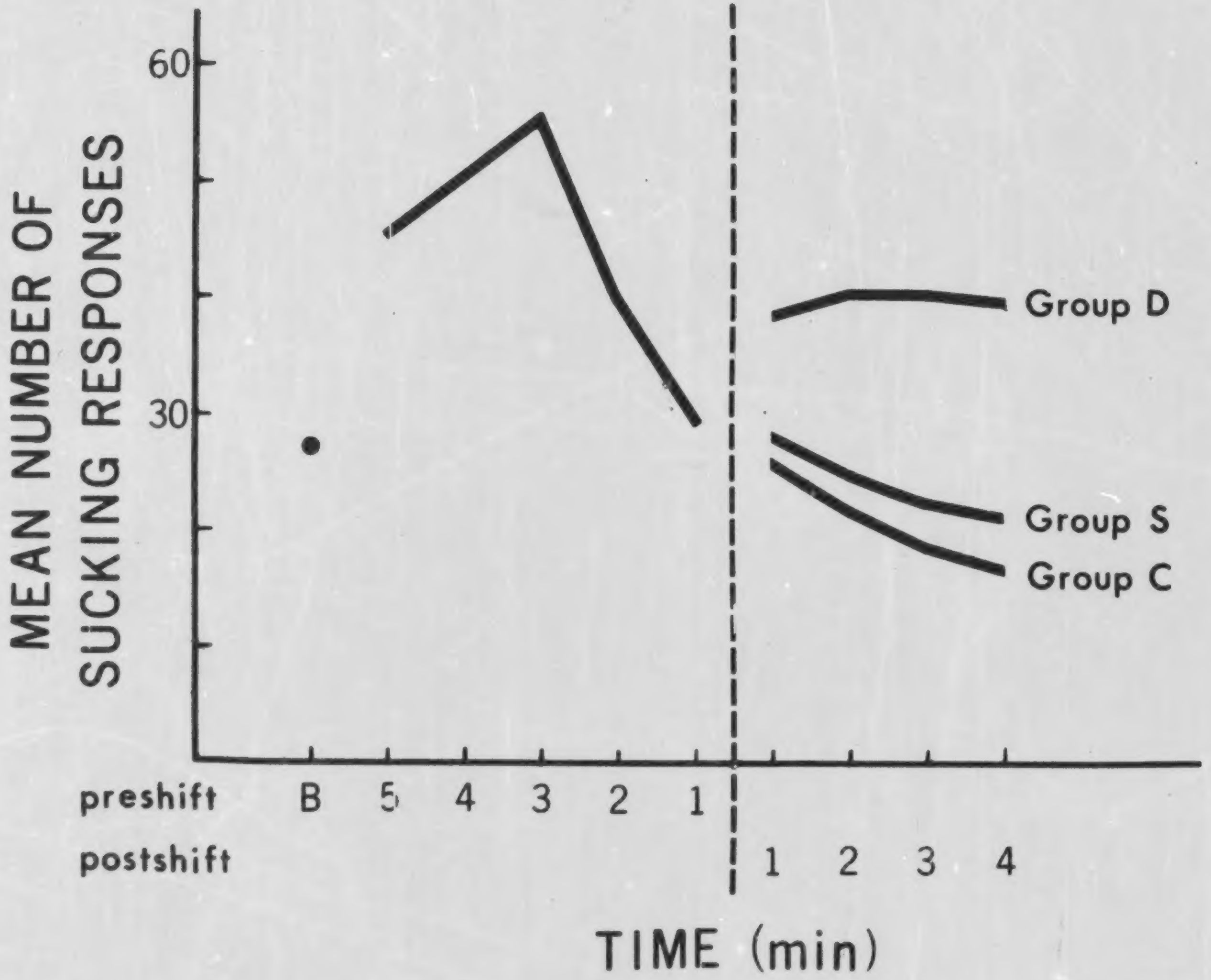


Figure 2: A schematic representation of the results of Eimas, Siqueland, Jusczyk, and Vigorito (1971) from tests in which infants were presented stimuli from a [ba]-to-[pa] continuum.

apparatus and to pull at the surrounding paraphernalia. The limitations set by infant age, however, appear to be rather minor. With respect to the younger infants, in particular, there is no reason to believe that they have learned much, if anything, about their to-be-native language between birth and four weeks.

Other problems on a more theoretical level are not so easily dismissed. For example, it happens that the phoneme boundary at about +25 msec VOT is only one of two such boundaries. An additional boundary occurs in many languages at about -30 to -50 msec VOT. One language which uses this phonetic boundary is Thai; English, of course, does not. If, indeed, the results of the first study are explainable in terms of innate phonetic feature analyzers, we would expect that there would be another set of innate analyzers tuned for stimuli with VOT values greater than -30 to -50 msec. These phonetic detection devices would not be needed in English, and perhaps in the English-speaking adult they have become inoperative from lack of use. Nevertheless, all infants in all cultures might be expected to be born with such a set of phonetic analyzers.

This notion was examined in a paradigm identical to that of the previous study, except here we looked at the infant's perception of the "Thai" boundary. D-pair stimuli had -70 and +10 msec values on the VOT continuum. S-pair stimuli had values of -150 and -70 msec. Spectrograms of the -150 msec and +10 msec stimuli are shown in Figure 1. In this experimental situation Group D infants yield a response rate pattern similar to that of all infants who listen to cross-boundary pairs; their postshift response rate is significantly greater than their preshift response rate. Group S infants, however, show a somewhat atypical result. They too yield a postshift increase, although the increase is slight and nonsignificant. Nevertheless, the difference between the two groups is not significant; and thus there is only a soupçon of evidence for innate phonetic feature detectors in this range of the VOT continuum. Thus far, the "Thai" boundary results are neither convincing nor embarrassing for the phonetic feature detector hypothesis.

Another small theoretical hurdle arises when we reconsider the results of the initial experiment in light of what is known about the Spanish voiced and voiceless stop consonants. The Spanish VOT boundary for labials is at +15 msec (Abramson and Lisker, 1973). This fact is important here because in Spanish, the D-pair stimuli of the initial study (+20 and +40 msec VOT) both lie within the [p] phoneme category. Would Spanish infants perceive these stimuli differently than American infants? Not likely. It appears that, in terms of VOT boundary values, English is a much more reasonable language than Spanish; or, at least, it has a phonetic boundary which conforms to that of more languages than does Spanish. Since the English boundary value is about +25 msec VOT and the Spanish boundary differs from it by only 10 msec, the differences between the phonetic perceptions of labials in the two languages might be accounted for by perceptual tuning that occurs over time, shifting the boundary slightly according to the constraints of the culture that the individual is reared in.

The final and most serious theoretical problem concerns the nature of voice onset time as a true continuum. It takes only a brief glance at Figure 1 to see that, as stimuli vary in VOT from -150 to +100 msec, more acoustic change occurs within certain time domains than within others. For example, there is little difference between stimuli of -150 and 0 msec VOT. All that separates them is a low-amplitude, low-frequency voice bar that barely registers on the amplitude display above the topmost spectrogram in Figure 1. Somewhat more acoustic change is manifested by differences in VOT from +40 to +100 msec as the first formant is cut back and the upper formants become aspirated. Nevertheless, the most prominent acoustic changes occur near the middle of this continuum between 0 and +40 msec VOT and it is here where the phonetic boundary lies. Along with changes in excitation of the upper formants, the first-formant transition is trimmed away piece-by-piece as stimuli increase in VOT, until at +40 msec there is no F1 transition. Imagine that, instead of phonetic feature detectors, there are auditory feature detectors that are triggered by low-frequency, rapidly rising frequency information[4] (see Whitfield and Evans, 1965). A +20 msec VOT stimulus has a brief transition, somewhat similar to that in -40, -20, or 0 msec VOT stimuli. A +40 msec VOT stimulus, on the other hand, has no transition and is seemingly more like +60, +80, and +100 msec stimuli. Thus, there appear to be two categories of stimuli: those with and those without F1 transitions. Although this explanation cannot account for the adult data on categorical perception (consider, for example, the Spanish VOT boundary of +15 msec), Stevens and Klatt (1972) have suggested that it might account for the infant data in the initial study.

Herein lies an issue of major theoretical importance: do neonates come equipped with phonetic feature detectors or with speech-relevant auditory detectors which are later incorporated into the language system? Although the voice-voiceless distinction may be the most important phonetic distinction in all languages, it cannot easily be used to settle this issue. A phonetic dimension which is purer in an acoustic sense must be used. Place would appear to be such a feature, particularly for stop consonants before front vowels.

Another phonetic feature: place. We, along with others (Moffitt, 1971; Morse, 1972), have investigated the infant's perception of stop consonants which differ only in place of articulation. For the sake of generality we selected another vowel, [æ] as in battle. For the sake of simplicity of discussion we will consider here an experiment which used two-formant stimuli, [bæ] and [dæ]. Mattingly et al. (1971) used these exact stimuli in a previous speech perception experiment with adult subjects. Schematic representations of six stimuli selected for the present study are shown in Figure 3. All were 245 msec in duration, with 15 msec of prerelease voicing, 40 msec of formant transitions, and 190 msec of steady-state vowel. Stimuli differed only in the trajectory of the F2 transition: Stimulus 1, [bæ], had an F2 transition which increased in frequency from a value of 1232 Hz to 1620 Hz, while at the other extreme Stimulus 6, [dæ], had a steady-state F2 of 1620 Hz. Equal increments of change in the initial frequency of the F2 transitions arrayed the six stimuli along an acoustic continuum. Since the [bæ]-[dæ] phoneme boundary occurs with a transition beginning at about 1500 Hz, Stimuli 1 and 4 were an S pair, while 2 and 5, and 3 and 6 were D pairs. Members of each pair differed by 230 Hz in the initial frequency of the F2 transition.

The experimental situation was the same as described earlier, and essentially so were the results.[5] As usual, Group C and S infants continue to decrease

---

[4] Here the term auditory detector is not meant to imply a peripheral mechanism. Instead, it refers to a more central mechanism that may be employed during linguistic and nonlinguistic processing alike.

[5] The ages of the infants in this study were two and three months.

FIGURE 3

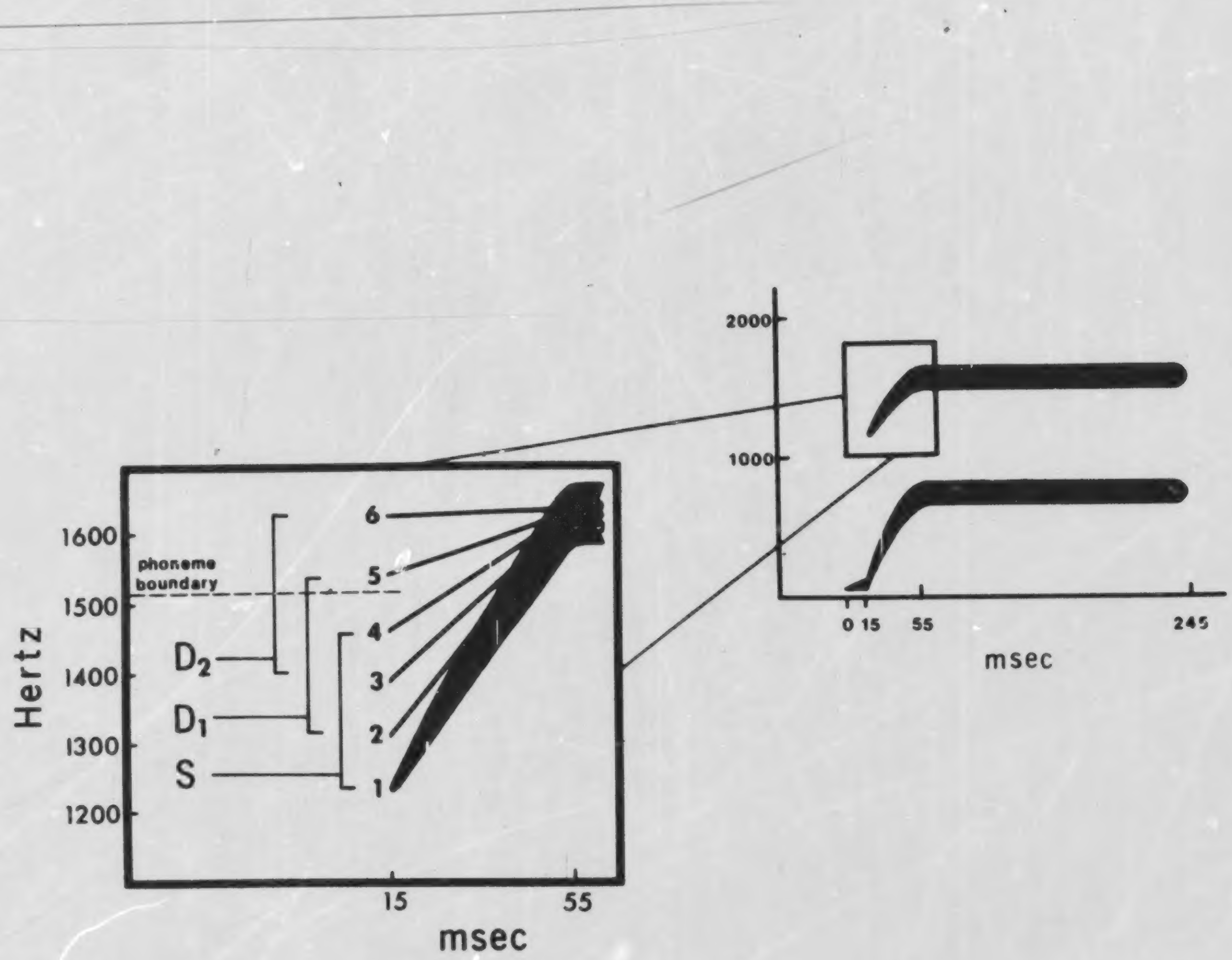


Figure 3: A schematic spectrogram of a two-formant [bæ] and the acoustic variation in the second-formant transition requisite for a [bæ]-to-[dæ] continuum.

their response rate in the postshift phase just as shown in Figure 2, while Group D infants increase their response rate with the advent of a stimulus beginning with a different phoneme. In this study, there was no difference between results of the D_1 and D_2 pairs. There was, however, a minor but interesting perturbation in the parallel between these results and those of the initial study represented in Figure 2. The postshift response function of the Group D infants differed slightly in shape. In Figure 2 the response rate for the first two minutes after the stimulus shift is approximately equal to that of the third and fourth minutes. In the present study this initial response rate was less than that of the third minute. However, in two other studies, the difference was even more pronounced and held for both the third and fourth minutes as well. In other words, the Group D infants in the place of articulation studies attained a postshift response peak later than those in the initial VOT study. Enough data have been collected to consider this a real and significant difference. It is interesting, then, to consider the place distinction in light of what we already know about the voicing distinction. Miller and Nicely (1955) have shown that, in listening to syllable-initial consonant phonemes under conditions of white noise, voicing is preserved better than place. Also Shankweiler and Studdert-Kennedy (1967) found that voicing was a more prominent feature than place in the lateralization of speech as measured by the results of a dichotic listening task. We seem to have found a functional parallel to the results of the adult studies along a dimension which might be termed phonetic-feature strength.

Again these results suggest that infants come into the world with bundles of phonetic feature analyzers. However, as with voicing, it remains possible that these discriminations could be made on the basis of auditory features alone. It happens that [dæ] syllables such as Stimuli 5 and 6 have very little, or no, second-formant transition, whereas [bæ] syllables such as Stimuli 1 through 4 have considerable F2 transitions. If infants had one auditory detector tuned to about 1500 Hz which was triggered by rapidly moving frequencies and another detector in the same range triggered by relatively constant frequencies, the results of the present study might easily be obtained.

Thus, in the present study we felt it necessary to include an additional experimental condition which considered this possibility. Instead of the entire speech stimulus, the F2 transition--the acoustic cue important for the phonetic categorization of the stop consonants--was excised and presented by itself. Mattingly et al. (1971) have named these stimuli "chirps" because of their resemblance to the discrete elements of birdsong (see, for example, Marler and Mundinger, 1971). Mattingly and his co-workers found that adults perceive these chirps differently in isolation than when they are part of a speech context. Would infants do the same? Yes, we found that, like adults, they perceive them differently.

The paradigm was again identical to that of previous studies, but the stimuli were S-chirp and D-chirp pairs as represented in the inset of Figure 3, not the entire S and D pair stimuli. The results showed that infants increase their response rate during the postshift phase for both types of stimulus pairs.

The overall results of both the chirp and the full two-formant stimulus conditions of this study are represented in Table 1. The pattern is remarkably similar to that for adults found by Mattingly et al. (1971). When the acoustic cue for a particular phoneme is part of the sound pattern of a speech utterance, infants can discriminate between items that lie across a phoneme boundary, but

TABLE 1: The perception of variation in the cue for place of articulation.

	Can the stimuli be discriminated?	
Speech cue	D pairs	S pairs
Within the speech context	yes	no
Alone as nonspeech	yes	yes

they cannot discriminate between items from within the same phoneme category. However, when these acoustic cues are isolated and presented by themselves, all pairs become about equally discriminable. Such a pattern, whether in infants or adults, suggests that the isolated chirps are perceived in what can be described as a nonspeech mode, whereas the whole speech utterance (which necessarily includes the F2 transition) is perceived in a speech mode (Mattingly et al., 1971).

Auditory feature analyzers. Do these results invalidate the notion that auditory feature analyzers (or in a narrower sense, detectors) are involved in the infants' perceptions? Not at all. The question arises, however, as to how they contribute to the perception of speech.

Whitfield and Evans (1965) demonstrated that certain cortical neurons in the cat respond vigorously to glissandi, but not to steady-state tones. They also found single neurons which responded to specific directions of frequency change. For example, some units responded to frequency changes from low-to-high, but not for the reverse direction. Moreover, this type of detector can be located in lower-than-cortical-level areas as well. Nelson, Erulkar, and Bryan (1966), for example, found frequency-change detectors at the level of the superior colliculus. We would expect that such auditory analyzers occur in humans as well.

Brady, House, and Stevens (1961) investigated adult human perception of sounds characterized by rapidly changing resonant frequencies, such as those found in the formant transitions of speech stimuli. Their results, like those of Mattingly et al. (1971), suggest that these nonspeech sound patterns are perceived differently than speech. However, the fact that they may be perceived in a different manner does not preclude the possibility that certain stages of auditory processing underlie speech processing. Consider the following example: Nabelek and Hirsh (1967) and Pollack (1967) found what could be described as two perceptual modes in the processing of frequency transitions in tone-like stimuli. The rules that govern the perception of long glissandi (those which last up to a second or more) are different from those rules for brief glissandi (less than 100 msec). More specifically, brief glissandi are more discriminable than longer ones with the same end-point frequencies. Interestingly, the brief glissandi most resemble formant transitions in speech, and we found that two- and three-month infants could perceive and discriminate these brief glissandi out of a speech context. Perhaps there are special auditory mechanisms for perceiving brief and rapidly changing auditory events, and these mechanisms contribute to speech perception by extracting from the acoustic signal speech-relevant information which can be used later in making a phonetic decision.

Cutting (in press) has elaborated this notion that there may be an intimate relation between the processing of speech and certain nonspeech sounds. It appears that transitional frequency information in complex nonspeech sounds is perceived in a manner similar to that of formant transitions in speech sounds. This inference stems from a result showing that, in dichotic listening, equal amounts of change in the transitions of both speech and nonspeech sounds appear to yield equal increments of perceptual change as measured by the right-ear advantage. In other words, adding formant transitions to vowel stimuli increases the magnitude of the right-ear advantage, and does so regardless of whether or not the transitions correspond to particular phoneme segments. Likewise, adding initial transitions to complex tone stimuli alters the resulting ear advantage in favor of the right-ear stimulus. Both results reflect an increase in the engagement of the processing mechanisms of the left-hemisphere system. Thus, the locus of speech-relevant auditory processing may be intimately related to the locus of speech perception. It is clear, however, that phonetic decisions are not merely the end result of auditory processing (see also Liberman et al., 1967).

We have seen that the auditory feature analyzers by themselves cannot accommodate the infants' perception of speech. Perhaps, auditory and phonetic feature analyzers function in a hierarchical manner to determine in a direct, sequential manner the results of the infant, and for that matter the adult, studies of speech perception. Recent evidence, however, suggests that the contribution of the auditory analysis is not very directly involved in the infant's discrimination of speech sounds. Eimas (in press) has found that infants are insensitive to differences in VOT when D pairs of to-be-discriminated stimuli differ in the magnitude of their voice onset difference. Whether the cross-boundary VOT difference is 20, 60, or even 100 msec, the relative increment in response rate does not change. Thus, stepping across the phoneme boundary is quantal and complete for the infant: there is no additive contribution of auditory analysis beyond that which is necessary for a phonetic decision. In this experimental situation auditory analysis appears to be too far removed from the phonetic decision-making process to affect the results. The same conclusion appears to be true for the processing of cues for place of articulation. Thus, we conclude that, although it is necessary that some form of auditory analysis extract the speech-relevant information from the acoustic signal, this information merely provides the input to the special speech processing analyzers. It is not necessarily an immediate antecedent of quantal phonetic decisions.

Phonetic feature analyzers. A number of researchers have begun to consider and explore models of speech perception that are based on feature detectors and do not require knowledge of production processes (see, for example, Lieberman, 1970; Abbs and Sussman, 1971; Stevens, 1972; Cole and Scott, 1972; Eimas, Cooper, and Corbit, 1973; Eimas and Corbit, 1973; Cooper, in press; Cutting, in press). Although these detector systems could be auditory or linguistic in nature, or perhaps both, we conclude that they must be primarily linguistic; that is, speech perception is mediated by phonetic feature analyzers that are sensitive to relatively restricted ranges of complex acoustic energy.

There are numerous problems that remain unresolved in hypothesizing a phonetic feature detector model. For example, what number and how many kinds of detectors are needed, and what is the nature of the invariant acoustic information? Nevertheless, a model of this kind accommodates much of the data and, moreover, does so by means of mechanisms that are analogous to the detector systems known to exist for the processing of complex visual information in man (McCollough, 1965; Blakemore and Campbell, 1969). The independent evidence for phonetic feature detectors is considerably less extensive than the evidence for visual detector systems. However, it is sufficient, we believe, to permit the inference that linguistic feature detectors exist and are the sole explanation for the results of the infant studies presented here.

In a recent series of experiments with adults, Eimas and Corbit (1973) obtained results that favor the existence of two phonetic feature detectors, one for the acoustic consequences of each of the two modes of voicing found in English and many other languages (Lisker and Abramson, 1964). They used a selective adaptation procedure in which the voiced or voiceless member of a phonetic contrast, such as [b]-[p], could be adapted by the repeated presentation of a good exemplar of that voicing mode. To measure the effects of adaptation, identification functions for bilabial and apical series of synthetic speech sounds, each of which differed only in VOT, were obtained from the same listeners in both an adapted state and an unadapted state. If detectors existed for two voicing distinctions, it was reasoned, then repeated presentation of a particular phoneme exemplar would fatigue the detector underlying its analysis and reduce that detector's sensitivity. As a consequence, the identification functions for the series of synthetic speech sounds would be altered. The results confirmed our expectations. After adaptation with a voiceless stop, either [p] or [t], listeners assigned fewer stimuli to the voiceless category, especially those stimuli near the original phonetic boundary. Adaptation with a voiced stop, either [b] or [d], had the opposite effect--fewer stimuli were heard as voiced stops. Adaptation, in essence, resulted in a shift in the locus of the phonetic boundary toward the adapting stimulus. It is particularly important to note that the effects of adaptation were very nearly the same whether or not the adapting stimulus and the to-be-identified stimulus were from the same series of speech sounds. That is, adaptation with the voiced stop [d] altered the phonetic boundary for an array of bilabial stops as effectively as did the bilabial stop [b]. It would appear, then, that the major effect of adaptation is to lower the sensitivity of the common voicing detector underlying the adapting stimulus and the members of the identification series.

Arguments are possible that the effects of adaptation are not sensory in nature, but rather reflect alterations in response decision factors. However, such explanations are difficult to defend given the fact that adaptation works across phoneme classes. Just how a response bias developed by the repeated presentation of [b], for example, might affect the tendency to assign the labels [d] or [t] to stimuli is not readily apparent. In addition, Sawusch and Pisoni (in press) have presented evidence that the identification functions for stop consonants are virtually unaffected by experimental manipulations that would be expected to produce marked changes in the assignment of phoneme labels according to the assumptions of both signal detection theory and adaptation-level theory. Such manipulations, however, do affect the identification of pure tones in a manner predicted by adaptation-level theory. (For other accounts of selective adaptation, see Ades, 1973; Bailey, 1973.)

In a second experiment, Eimas and Corbit (1973) showed that selective adaptation could also alter the locus of the peak of discriminability in a series of bilabial stop consonants. This evidence strongly indicates that the discriminability of such an array of synthetic stops is based on the manner in which acoustic information is assigned to phonetic feature categories. Alterations in phonetic decision criteria, as measured by a shift in the locus of the phonetic boundary, were matched almost perfectly by the shift in the locus of the discriminability peak. Furthermore, Eimas, Cooper, and Corbit (1973) found that the size of adaptation was central and specific to the speech processing system. Presentation of the adapting stimulus to one ear and the identification series to the other, unadapted ear did not alter either the direction or magnitude of the adaptation effects. However, when the voicing information, used for adaptation, was presented in a nonspeech context, there were no reliable or systematic effects of adaptation. This was true despite the exhortations of the experimenters suggesting to some subjects that they might try to perceive the adapting stimuli as speech.

Finally, Cooper (in press) has obtained evidence for the existence of phonetic feature detectors that mediate the perception of the three major distinctions for place of articulation. Using a selective adaptation procedure again and a series of synthetic speech sounds that varied in the starting frequency and direction of the second- and third-formant transitions, Cooper found marked shifts in the loci of the phonetic boundaries and peaks in the discriminability functions. All of these shifts were consistent with the assumption of three independent feature detectors.

A model of speech perception based on phonetic feature detectors yields a number of advantages as well as being able to accommodate much of the data on the perception of segmental units by both infants and adults. Eimas and Corbit (1973) have outlined a feature detector model that can account for the adult discrimination and identification data with and without adaptation. Eimas (1973, in press) has extended the analysis to explain the data from infant studies of speech perception. It is with this extension that we will now be concerned. To explain the infant's ability to discriminate variations in the cues for voicing and place of articulation by reference to phonetic feature values, we need first to assume the presence of appropriate phonetic feature analyzers. These analyzers, by inference from our infant studies, must be operative shortly after birth, perhaps having been set in operation merely by experiencing speech. Given the passive nature of a feature detector analysis, the presentation of a signal with sufficient linguistic information to activate the speech processing mechanisms will excite each of the phonetic feature analyzers for which there is an adequate stimulus. The repeated presentation of the same stimulus, which occurs in the infant studies, will result in the adaptation of the activated detectors (see also Eimas and Corbit, 1973). Adaptation of the detectors, which presumably results in the diminution of their output signals, may well be related to the decrement in the reinforcing properties of novel stimuli and the subsequent decrement in the infant's response rate. The presentation of a second speech stimulus, which, although acoustically different, excites the same set of detectors, will not be experienced as a novel stimulus by the infant. Consequently, there will be no increased effort to obtain this stimulus. Introduction of a second stimulus that activates one or more different detectors, on the other hand, yields a different set of phonetic feature values and will be experienced as novel. Our notion is that the infant increases his response rate in order to obtain this new perception. From phonetic-feature detectability and from the infant's well-established appetite for novel stimulation, it can easily be predicted that infants in Groups S and C will, on the average, show continued decrement in the conditioned response rate during the final four minutes of the experiment. The infants in Group D, by contrast, will show a marked increase in the rate of response during the postshift minutes. This increment is unrelated to the amount of acoustic difference between the two stimuli. To explain the continuous discrimination of variations in second-formant transitions in a nonspeech context, we need to assume that there was not sufficient linguistic information in the stimulus to activate the speech processing mechanisms and that the processing of these sounds was limited to the more general, auditory mechanisms (see Eimas et al., 1973; Mattingly et al., 1971).

We should consider why infants (and adults) are able to discriminate within-category variations of the second-formant transition in nonspeech settings, and yet are unable to discriminate the same information in speech contexts. It seems reasonable to assume that an acoustic event, whether speech or nonspeech, undergoes much of the same auditory processing (see Cutting, in press). Hence, the failure to discriminate the same information in one context presents something of a paradox. Perhaps it is simply the case that the output of speech processing mechanisms takes precedence over the output of nonspeech auditory mechanisms. Or it is possible that phonetic processing requires more time than auditory analysis, and that at the conclusion of phonetic feature extraction, the relatively brief but complex auditory information, which signals consonantal features, has faded or in some other manner become unavailable to the response decision component (Fujisaki and Kawashima, 1968; Liberman, Mattingly, and Turvey, 1972). The evidence to date does not permit us to choose among these and other explanations.

With regard to the advantages of linguistic detectors, the analysis of speech into phonetic features requires only that the acoustic event has sufficient linguistic information (phonetic information, for our purpose) to activate the speech processing mechanisms, and that the detectors be present and operative. This form of analysis does not require a decision on the listener's part that the acoustic signal is speech and, hence, in need of special processing. Analysis-by-feature-detection is an automatic and passive process, and as such provides the infant with the means for the immediate recognition of speech and the means for parsing speech into discrete elements. These factors must surely hasten the acquisition of speech. Were it necessary for infants to learn that speech requires special processing and that speech is composed of discrete elements despite its continuous form, the acquisition of language would be a difficult and tedious process, if indeed language could be learned at all. Finally, the automatic analysis of speech into distinctive and invariant features provides the infant with a set of anchor points by which he can eventually come to recognize and master the considerable amount of context-conditioned variation found at all levels of language. That some of these anchor points might be more innate than others is not a serious problem. Without some degree of invariance in both the signal and processes of analysis, the acquisition of human languages would not be possible in the relatively short time that it takes a child to become a proficient user of language.

REFERENCES

Abbs, J. H. and H. M. Sussman. (1971) Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. J. Speech Hearing Res. 14, 23-36.

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In Proceedings of the Fifth International Congress of Acoustics, A-51, Liege.

Abramson, A. S. and L. Lisker. (1970) Discriminability along the voicing continuum: Cross-language tests. In Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967 (Prague: Academia) 15-25.

Abramson, A. S. and L. Lisker. (1973) Voice-timing perception in Spanish word-initial stops. J. Phonetics 1, 1-8.

Ades, A. E. (1973) Some effects of adaptation on speech perception. Quarterly Progress Report (Research Laboratory of Electronics, MIT) *111*, 121-129.

Bailey, P. J. (1973) Perceptual adaptation for acoustical features in speech. Speech Perception, Report on Research in Progress (Queen's University, Belfast, Northern Ireland) *2*, 29-34.

Bernard, J. and L. W. Sontag. (1947) Fetal reactivity to tonal stimulation: A preliminary report. J. Genet. Psychol. *70*, 205-210.

Blakemore, C. and F. W. Campbell. (1969) On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. J. Physiol. *203*, 237-260.

Bloom, L. (1970) *Language Development: Form and Function in Emerging Grammars*. (Cambridge, Mass.: MIT Press).

Brady, P. T., A. S. House, and K. N. Stevens. (1961) Perception of sounds characterized by a rapidly changing resonant frequency. J. Acoust. Soc. Amer. *33*, 1307-1362.

Brown, R. (1973) *A First Language; The Early Stages*. (Cambridge, Mass.: Harvard University Press).

Cole, R. A. and B. Scott. (1972) Phoneme feature detectors. Paper presented at the meeting of the Eastern Psychological Association, Boston, Mass., April.

Cooper, W. E. (in press) Adaptation of phonetic feature analyzers for place of articulation. J. Acoust. Soc. Amer.

Cutting, J. E. (in press) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys.

Dorman, M. F. (in press) Auditory evoked correlates of speech sound discrimination. Percept. Psychophys.

Eimas, P. D. (1973) Linguistic processing of speech by young infants. Paper presented at the conference "Language Intervention with the Mentally Retarded," Wisconsin Dells, Wis., June.

Eimas, P. D. (in press) Speech perception in early infancy. In *Infant Perception*, ed. by L. B. Cohen and P. Salapatek. (New York: Academic Press).

Eimas, P. D., W. E. Cooper, and J. D. Corbit. (1973) Some properties of linguistic feature detectors. Percept. Psychophys. *13*, 247-252.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. *4*, 99-109.

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. M. Vigorito. (1971) Speech perception in infants. Science *171*, 303-306.

Fujisaki, H. and T. Kawashima. (1968) The influence of various factors on the identification and discrimination of synthetic speech sounds. In *Reports of the Sixth International Congress on Acoustics*, Tokyo, August.

Graham, F. K. and J. C. Jackson. (1970) Arousal systems and infants' heart rate responses. In *Advances in Child Development and Behavior*, Vol. 5, ed. by H. E. Reese and L. P. Lipsitt. (New York: Academic Press) 59-117.

Jakobson, R. (1968) *Child Language, Aphasia, and Phonological Universals*, trans. by A. Keiler. (The Hague: Mouton).

Jakobson, R., C. G. M. Fant, and M. Halle. (1951) *Preliminaries to Speech Analysis*. (Cambridge, Mass.: MIT Press).

Kaplan, E. L. and G. Kaplan. (1971) The prelinguistic child. In *Human Development and Cognitive Processes*, ed. by J. Eliot. (New York: Holt, Rinehart, and Winston) 358-381.

Karmel, B. Z. (1969) The effect of age, complexity, and amount of contour on pattern preferences in human infants. J. Exp. Child Psychol. *7*, 339-354.

Kessen, W., M. M. Haith, and P. H. Salapatek. (1970) Infancy. In *Carmichael's Manual of Child Psychology*, 3rd ed., ed. by P. H. Mussen. (New York: Wiley) 287-445.

Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Amer. 29, 117-123.

Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., K. S. Harris, J. A. Kinney, and H. Lane. (1961) The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. 61, 379-388.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington, D.C.: V. H. Winston).

Lieberman, P. (1970) Towards a unified phonetic theory. Ling. Inq. 1, 307-322.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. 74, 287-307.

Lieberman, P., K. S. Harris, P. Wolff, and L. H. Russell. (1972) Newborn infant cry and nonhuman primate vocalization. J. Speech Hearing Res. 14, 719-727.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English stops. Lang. Speech 10, 1-28

Lisker, L. and A. S. Abramson. (1970) The voicing dimension: Some experiments in comparative phonetics. In Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967 (Prague: Academia).

Marler, P. and P. Mundinger. (1971) Vocal learning in birds. In Ontogeny of Vertebrate Behavior. (New York: Academic Press).

Mattingly, I. G. (1968) Synthesis by rule of General American English. Ph.D. dissertation, Yale University. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

McCollough, C. (1965) Color adaptation of edge-detectors in the human visual system. Science 149, 1115-1116.

McNeill, D. (1970) The Acquisition of Language. (New York: Harper & Row).

Menyuk, P. (1971) The Acquisition and Development of Language. (Englewood Cliffs, N. J.: Prentice-Hall).

Miller, G. A. and P. Nicely. (1955) An analysis of perception confusions among some English consonants. J. Acoust. Soc. Amer. 27, 338-352.

Moffitt, A. R. (1971) Consonant cue perception by twenty- to twenty-four-week-old infants. Child Development 42, 717-731.

Molfese, D. L. (1972) Cerebral asymmetry in infants, children, and adults: Auditory evoked responses to speech and noise stimuli. Unpublished Ph.D. dissertation, Pennsylvania State University (Psychology).

Morse, P. A. (1972) The discrimination of speech and nonspeech stimuli in early infancy. J. Exp. Child Psychol. 14, 477-492.

Nabelek, T. and I. J. Hirsh. (1967) On the discrimination of frequency transitions. J. Acoust. Soc. Amer. 45, 1510-1519.

Nelson, P. G., S. D. Erulkar, and S. S. Bryan. (1966) Response of units of the inferior-colliculus to time-varying acoustic stimuli. J. Neurophysiol. 29, 834-860.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. dissertation, University of Michigan (Psycholinguistics). (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.

Pollack, I. (1967) Detection of rate of change of auditory frequency. J. Exp. Psychol. 77, 535-554.

Sawusch, J. R. and D. B. Pisoni. (in press) Category boundaries for speech and nonspeech sounds. J. Acoust. Soc. Amer.

Shankweiler, D. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Siqueland, E. R. and C. A. DeLucia. (1969) Visual reinforcement of nonnutritive sucking in human infants. Science 165, 1144-1146.

Slobin, D. I. (1971) Psycholinguistics. (Glenview, Ill.: Scott Foresman).

Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In Human Communication: A Unified View, ed. by E. E. David, Jr. and P. B. Denes. (New York: McGraw-Hill) 51-66.

Stevens, K. N. and D. H. Klatt. (1972) The role of formant transitions in the voice-voiceless distinction for stops. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 101, 188-197.

Studdert-Kennedy, M. (in press) The perception of speech. In Current Trends in Linguistics, Vol XII, ed. by T. A. Sebeok. (The Hague: Mouton).

Wertheimer, M. (1961) Psychomotor coordination of auditory and visual space at birth. Science 134, 1962.

Whitfield, I. C. and E. F. Evans. (1965) Responses of auditory cortical neurons to stimuli of changing frequency. J. Neurophysiol. 28, 655-672.

Wolff, P. H. (1966) The natural history of crying and other vocalization in early infancy. In Determinants of Infant Behavior, Vol. IV, ed. by B. M. Foss. (London: Methuen) 81-109.

An Experimental Evaluation of the EMG Data Processing System: Time Constant Choice for Digital Integration

Diane Kewley-Port[+]
Haskins Laboratories, New Haven, Conn.

DEFINITION OF THE PROBLEM

The design of the Haskins Laboratories' electromyographic (EMG) data processing system was based on certain premises concerning the nature of EMG signals in relation to articulatory movement (Cooper, 1965; Port, 1971; Kewley-Port, 1973). We are currently attempting to demonstrate experimentally the validity of these premises. The first premise is that the time-varying signal observed at a pair of electrodes is the sum of the desired EMG signal and a noise signal. The desired EMG signal is defined as the energy summation over time from a population of firing motor units. The noise signal is considered to be statistically random (with a mean of zero) and arises from the phase differences among different motor units' potentials. To eliminate the noise and obtain the desired EMG signal, time-varying signals from many repetitions of the same utterance are aligned carefully in time, sampled, and averaged under computer control. This averaging eliminates the noise--since its mean is zero--and produces an average of the desired EMG signals.

In order to sample the time-varying signal, a second premise was made concerning the importance of the high frequency components (above 10,000 Hz) of the signal. It was assumed that the time variation of the EMG signals important for speech research would be about the same as the time change of articulatory movement, which is of the order of 20 msec or less. Accordingly, a sample rate of 200 Hz was chosen. This necessitated preprocessing of the signal, including rectification and hardware integration, before sampling. Until recently EMG preprocessing utilized standard RC integrators with a time constant of about 22 msec. These integrators were replaced with linear-reset integrators with a 5 msec time constant in September 1973. The linear-reset integrators provide essentially true time integration since the energy is summed over the 5 msec interval, sampled, and then reset to zero before the next 5 msec interval. Furthermore, Kreifeldt (1971) has shown that a linear function of integration is superior to that obtained from RC integrators for smoothing EMG signals.

Further digital smoothing of the sampled signal for comparison with articulatory events is at the discretion of the experimenter. Computer programs provide visual displays of the effects of increasing the smoothing in 5 msec increments, called the time constant of integration. The integration is both forward and backward by means of a linear weighting function.

[+] Also the Graduate Center, City University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

EXPERIMENTAL ANALYSIS

Several questions about the above premises and techniques can be examined experimentally by placing a number of electrodes bilaterally in a muscle expected to be functionally undifferentiated. That is, we wish to obtain EMG signals simultaneously at several electrode placements that are equally representative of the muscle's action pattern. Two experiments have been conducted, one using the levator palatini and one using the mylohyoid. Visual inspection of the data suggests that only the electrodes in the levator palatini produced reasonably equivalent EMG signals.

Using the data from these two experiments, correlation analysis enables us to examine several questions. How representative is one electrode placement of the EMG activity of the muscle as a whole? To what extent is the "noise" component of the EMG signal truly random? In what ways is an utterance spoken in a list the same as, or different from, that utterance spoken in running speech?

For two more questions, the analyses are completed. The first is, what time constant of integration should be chosen for the digital smoothing of the EMG signals from individual utterances prior to averaging? In general, the smallest time constant that will effectively smooth the EMG signals is desired. Among the parameters likely to influence the choice of time constant is the level of EMG activity picked up by the electrodes. It appears from visual inspection of the EMG signals that more smoothing is needed for signals with high levels of activity. Thus, two mylohyoid electrodes (labeled Channel 3 and Channel 6 in the following figures) with signal peaks around 500 to 600 microvolts were chosen. Other factors influencing the amount of smoothing needed include the kind and length of utterance. In this experiment, two kinds of utterances were used. A text was read and segments of sentences about 2 sec long were sampled from the text. Also, phrases appearing in the text were read in list form. Two sentences and two 1-sec phrases were selected for analysis.

The procedure was to run the computer programs that smooth the sampled EMG signals before averaging several times using different time constants. The sampled (unsmoothed) data has a base time constant of 5 msec from the linear-reset integrators in preprocessing. Time constants chosen for digital integration were 15, 25, 35, 45, 55, 65, and 95 msec. For each time constant, electrode channel, and utterance a correlation analysis was made. An example of the averaged EMG signal and eight (out of 14) of the signals going into the average appear in Figure 1 for Channel 3 with a time constant of five msec (i.e., no digital smoothing). With increased digital smoothing, the individual EMG signals begin to look more and more like the average signal, as can be seen in Figure 2 for the same data at a time constant of 95 msec. However, with increased smoothing the times of onset and offset of EMG activity become smeared and peaks of activity become broader and lower.

Usually it is desirable to minimize these effects of integration by choosing the smallest time constant that will produce a smooth looking average. To find this time constant, the Pearson product moment correlation coefficient r is computed between the samples of an individual EMG signal and its average. As the time constant increases, r will increase since both signals will have less ripple. To obtain a function representing this increase in r, ten individual EMG signals were correlated with their mutual average for each time constant, and an average of these correlation coefficients was computed. Figures 3 and 4 show this average

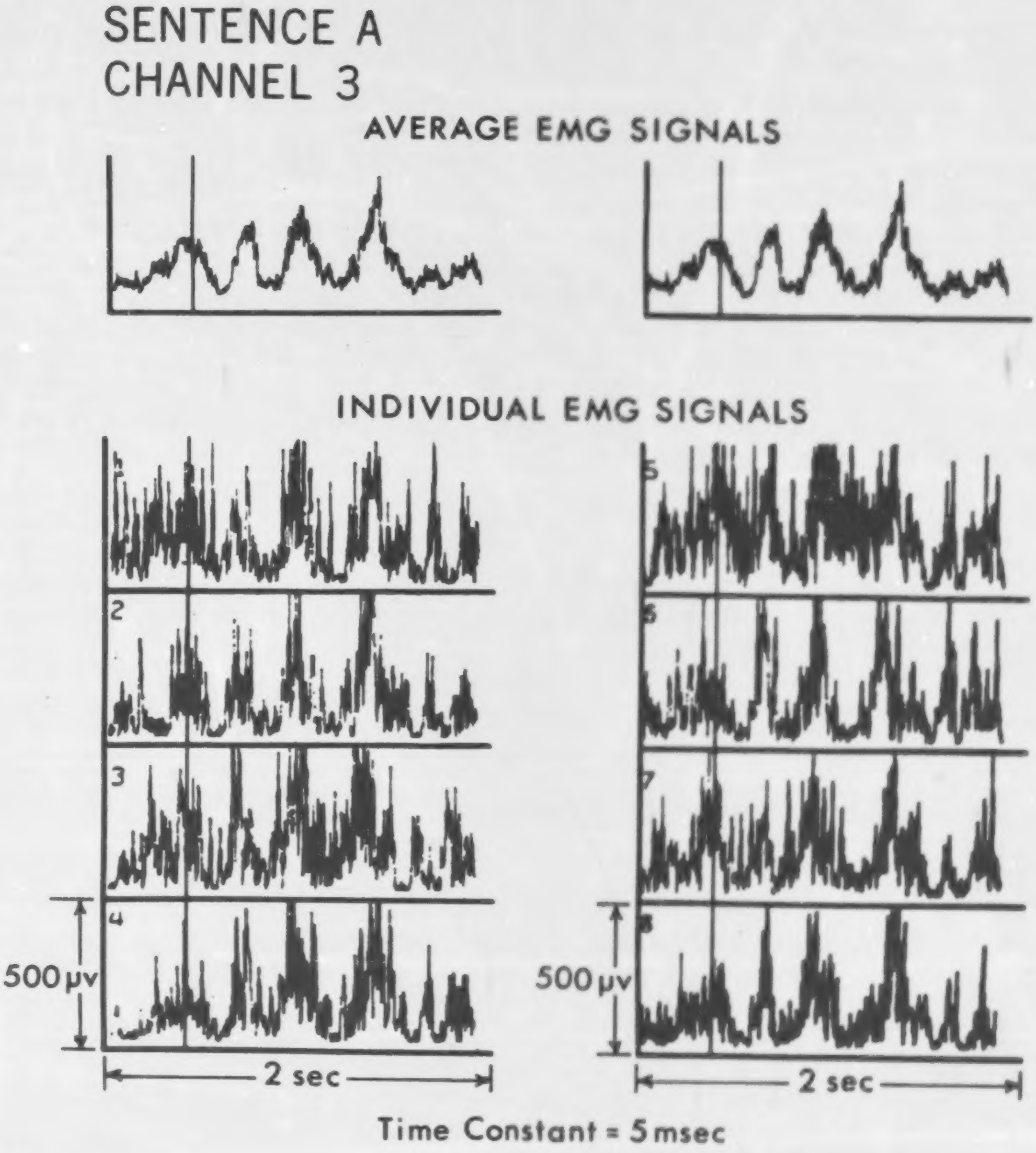


Figure 1: EMG signals from the mylohyoid as seen on the computer driven storage oscilloscope for the first 2 sec of the sentence "Eve and Clayton left Kansas for the...." The average EMG signal is for 14 repetitions of the sentence. The first 8 of 14 individual repetitions are displayed. The time constant is 5 msec (that is, these are samples obtained from the linear-reset integrators with no digital smoothing).

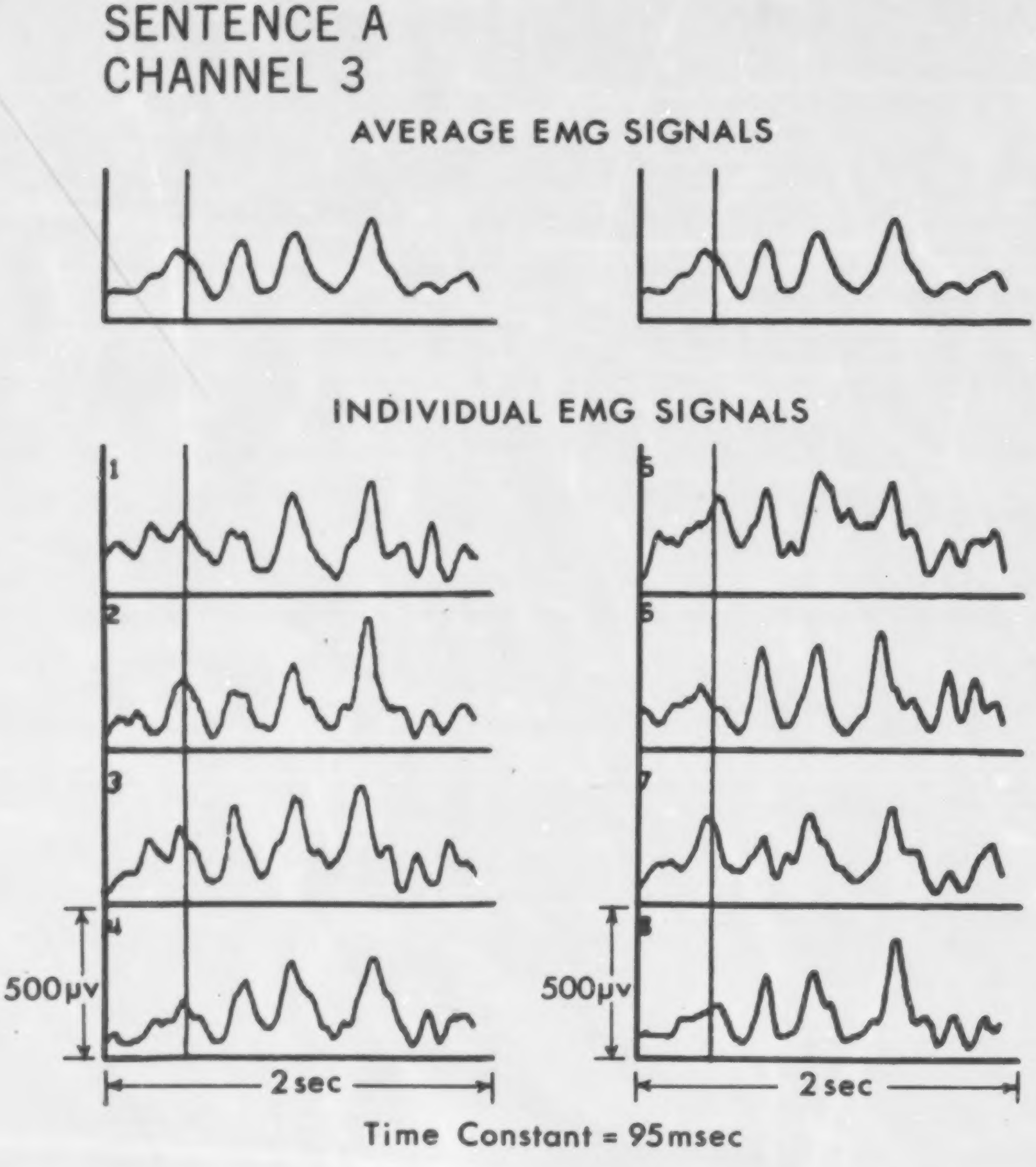


Figure 2: The same data as in Figure 1, but the EMG signals have been digitally smoothed with a time constant of 95 msec before averaging.

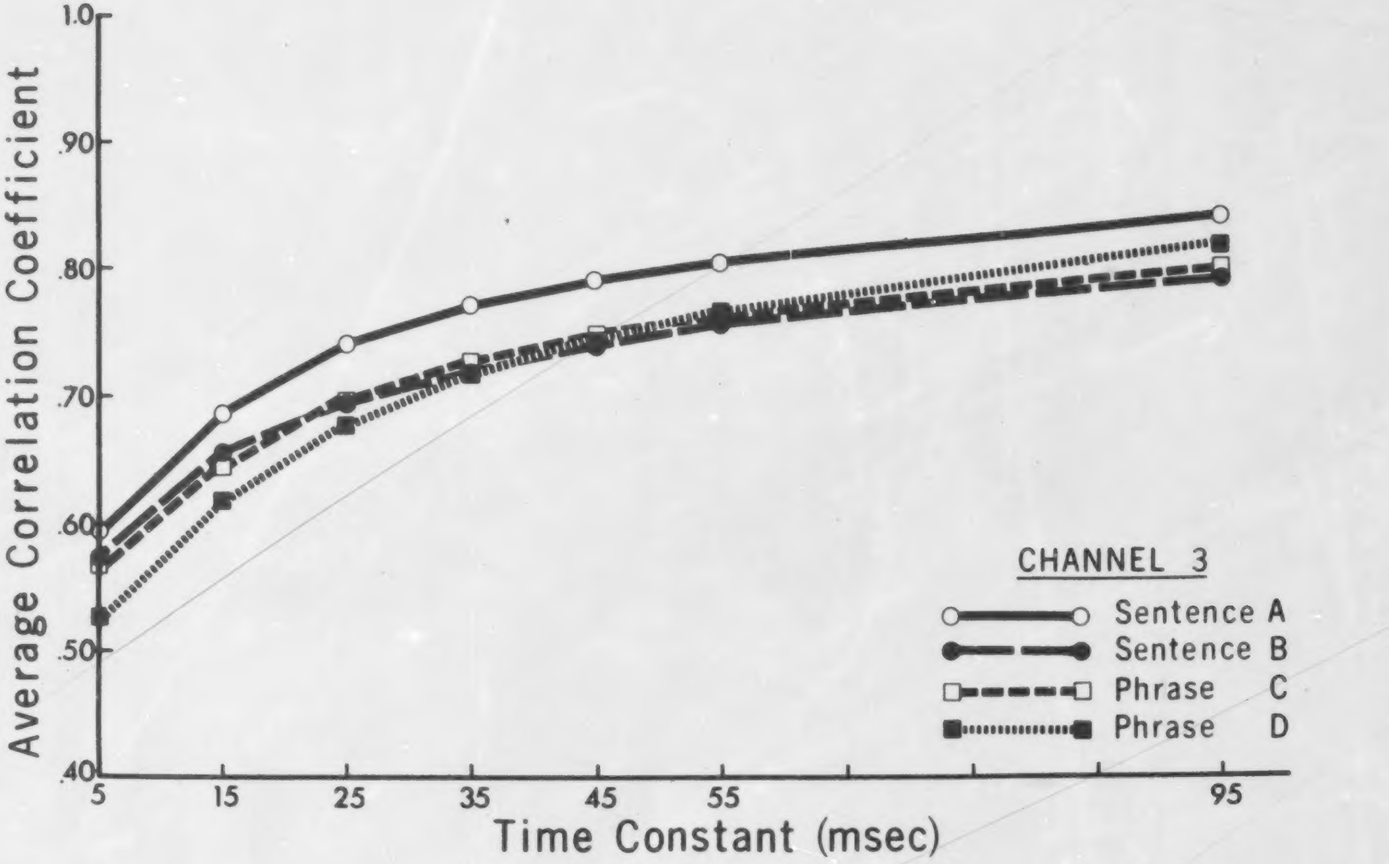


FIGURE 3

Figure 3: Each point is the average value of ten correlation coefficients computed between ten individual utterances and their average as the time constant of digital smoothing varies from 5 to 95 msec. The functions connect points for two sentences and two phrases on mylohyoid Channel 3.

FIGURE 4

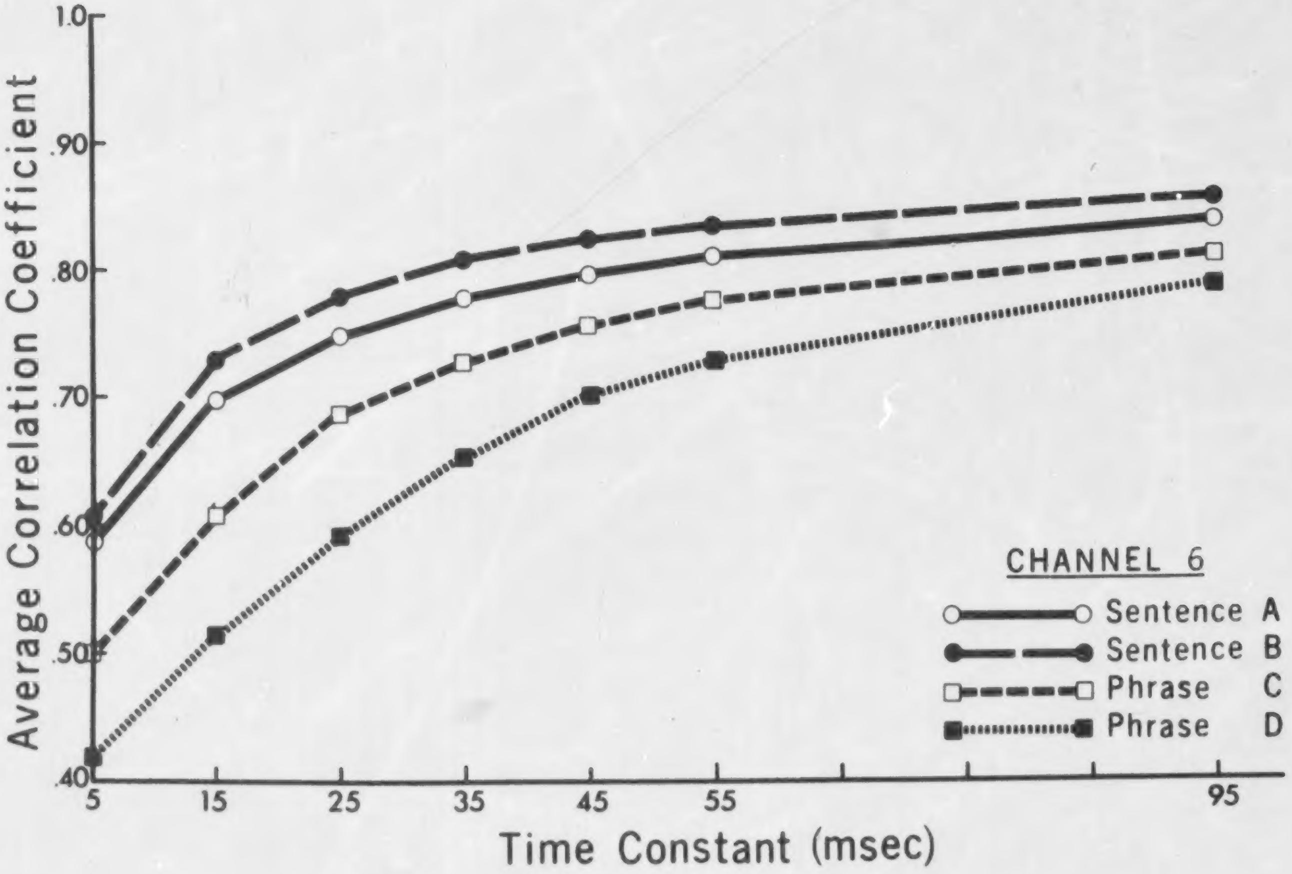


Figure 4: Each point is the average value of ten correlation coefficients computed between ten individual utterances and their average as the time constant of digital smoothing varies from 5 to 95 msec. The functions connect points for two sentences and two phrases on mylohyoid Channel 6.

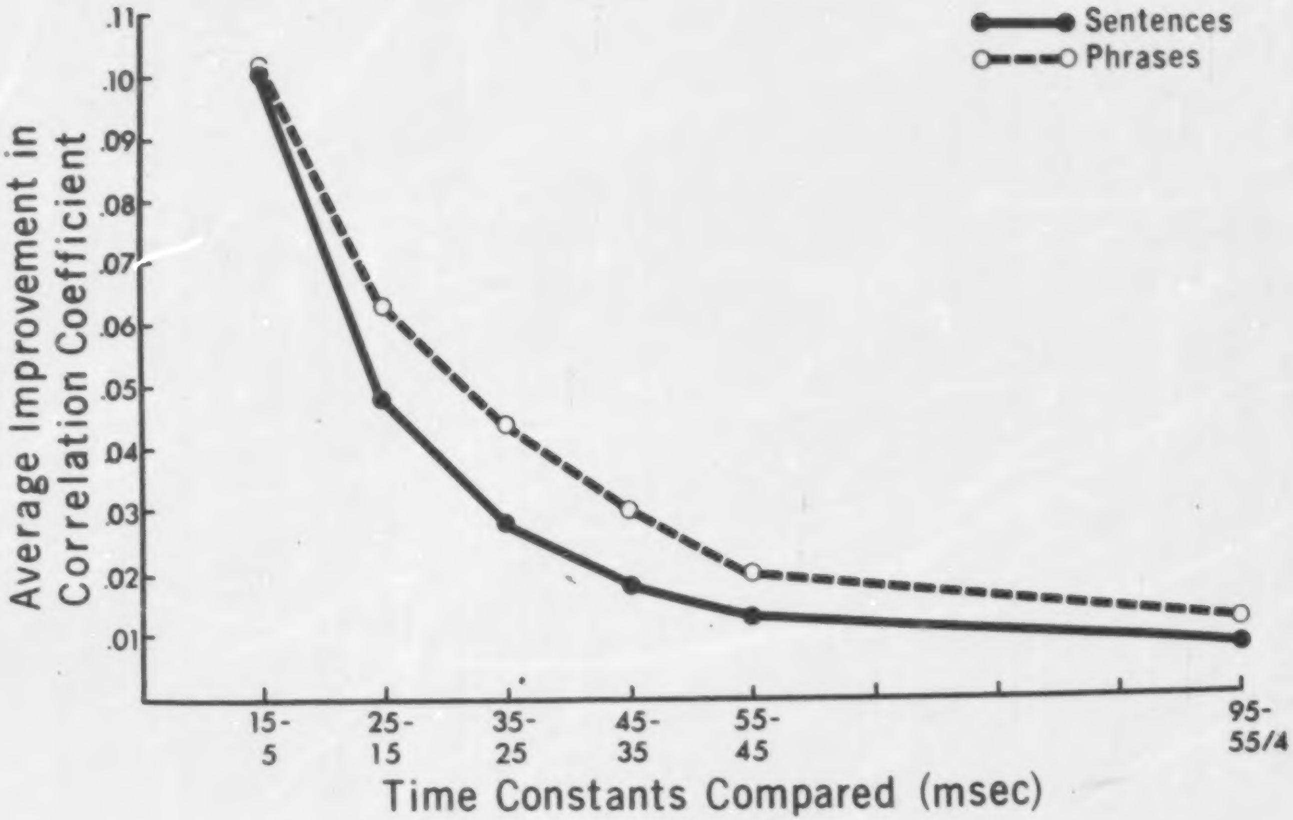


Figure 5: Each point is the difference in the correlation coefficients calculated for two time constants averaged over mylohyoid Channels 3 and 6. "Sentences" is an average for Sentences A and B; "Phrases" is an average for Phrases C and D. Time constants are varied in 10-msec intervals, except the last interval which is 40 msec. For comparison the difference in r's from 55 to 95 msec is divided by 4.

FIGURE 5

correlation coefficient function for two sentences and two phrases on two different mylohyoid electrode channels.

These figures show that average r increases rapidly when time constants are less than 25 msec. The functions do not reach an asymptote. Differences can be seen between Channels 3 and 6. Although the data were sampled from both channels simultaneously, Sentence A always had the highest average r for Channel 3, but Sentence B had the highest on Channel 6. There is more spread between the functions on Channel 6 than on Channel 3. We can see a tendency for higher average r's to occur in the sentences than in the phrases. This appears to be an effect of greater context constraint in the sentences during running speech.

All of the functions in Figures 3 and 4 appear to increase in a similar way. To check this, the average r's for each 10 msec increase in time constant were subtracted, giving functions of improvement in average r. The values obtained between Channels 3 and 6 were the same, but there were differences between the sentences and the phrases. In Figure 5 average improvement functions are plotted separately for sentences and phrases averaged over Channels 3 and 6. We can see that r is improved by only .03 for a time constant of 35 msec for sentences and for a time constant of 45 msec for phrases.

We conclude, then, that there is no best time constant for digital smoothing of the EMG signals. Figures 3, 4, and 5, however, should assist an experimenter in choosing a time constant. These data received little benefit in smoothing from time constants greater than 45 msec. The experimenter interested in peak height differences should probably choose a smaller time constant to minimize the peak lowering effects of smoothing mentioned before.

These data also are relevant to the second question: to what extent does averaging obscure phonetically significant variation in EMG signals between different repetitions of the same utterance? For the utterances and the channels examined, if the time constant chosen was 35 to 45 msec, the results show that average correlations fall between .70 and .80 (Figures 3 and 4). This is a quantitative indication of the extent to which the individual EMG signals have the same pattern of activity (covary) as the average signals. That is, these signals vary in the same direction at the same moment in time during 50 to 65 percent of all samples. We plan to analyze further the components of the observed variation, but the present analysis makes clear that signals of like kind are being averaged.

REFERENCES

Cooper, Franklin S. (1965) Research techniques and instrumentation: EMG. In Proceedings of the Conference: Communicative Problems in Cleft Palate, ASHA Reports No. 1, 153-168.

Kewley-Port, Diane. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-183.

Kreifeldt, John G. (1971) Signal versus noise characteristics of filtered EMG used as a control source. IEEE Trans. Biomedical Engineering BME-18, 16-22.

Port, Diane K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

More on the Motor Organization of Speech Gestures*

Fredericka Bell-Berti[+] and Katherine S. Harris[++]
Haskins Laboratories, New Haven, Conn.

We have reported before observations of a reorganization of motor commands to muscles whose increased contraction will further narrow some portion of the upper vocal tract (Bell-Berti and Harris, 1973). This reorganization manifests itself as the merging of electromyographic (EMG) activity for two contiguous speech gestures when the second gesture requires a more closed vocal tract than the first (for example, a vowel-consonant syllable), and as the maintenance of separate activity peaks when the second gesture requires a more open vocal tract than the first (for example, a consonant-vowel syllable).

Another statement of this hypothesis might be: when a muscle must be shorter for the second element in a sequence than for the first, the motor commands for the two gestures will merge into one; when a muscle must be lengthened for the second element of a sequence, activity will be suppressed between commands for the two gestures, and the two commands will not merge. In this paper we will extend our statement about muscles that are vocal tract closers.

Anticipatory coarticulation is a phenomenon that has been described in several situations: the lip-rounding of a vowel anticipated in a preceding string of consonants, or consonant nasality anticipated in preceding vowels, for example. Henke's (1966) "look-ahead" model of anticipatory coarticulation predicts that a feature will be anticipated as soon as it is not contradicted in the intervening speech string. We have been looking for instances of anticipatory coarticulation at the motor command level. Admittedly, this is a very different level than that at which most of the work on anticipatory coarticulation has been done.

We will begin by reexamining some data from Bell-Berti and Harris (1973) and considering how it might be interpreted in light of the Henke model.

Figure 1 shows examples of EMG activity from the genioglossus muscles of three speakers of American English repeating utterances having /-ik-/ and /-ki-/ sequences embedded in them. The genioglossus muscle raises and bunches the tongue for /i/ and /k/ segments, thus narrowing the vocal tract.

---

*Paper presented at the 87th meeting of the Acoustical Society of America, New York, April 1974.

[+] Also Montclair State College, Upper Montclair, N. J.

[++] Also the Graduate School and University Center of the City University of New York.

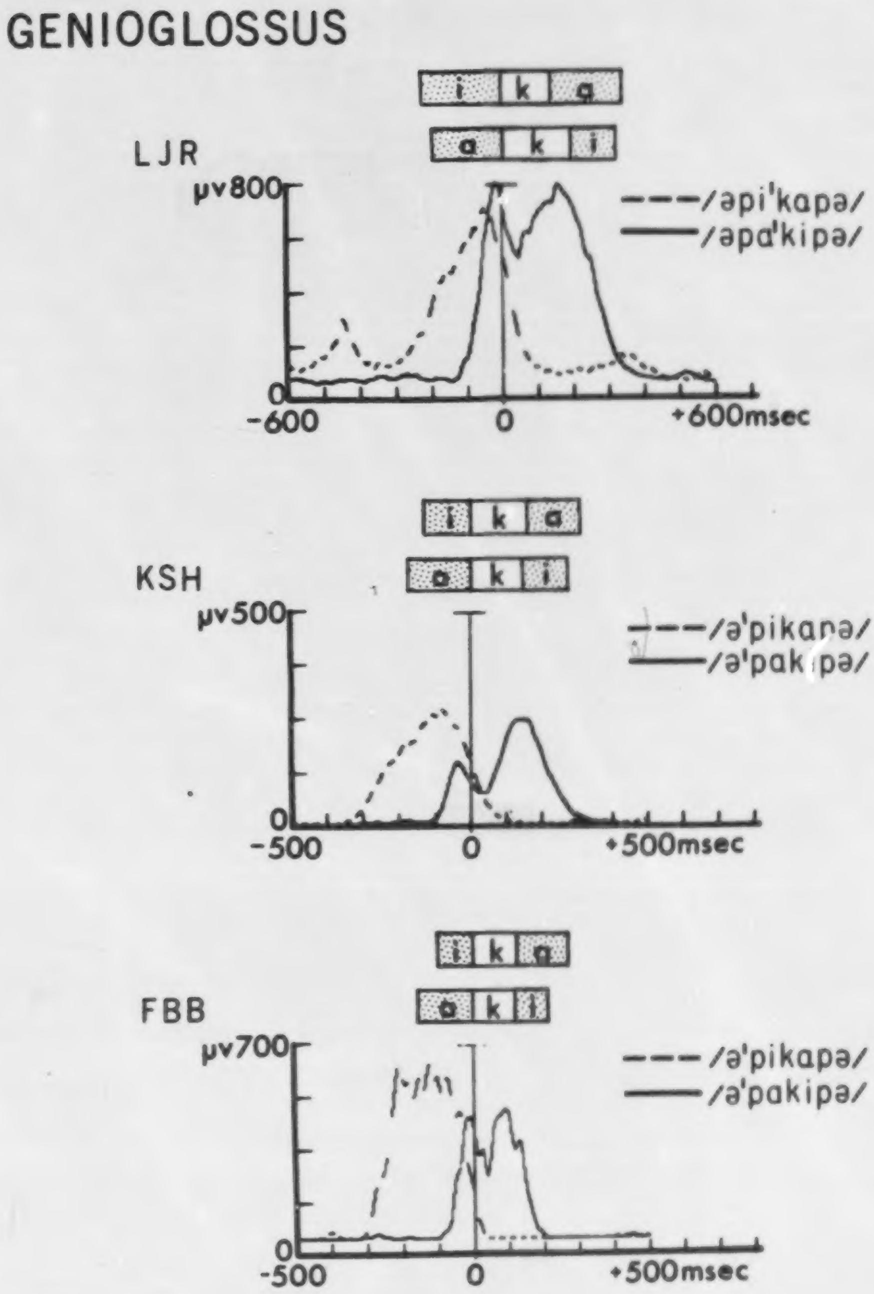

GENIOGLOSSUS
LJR
i k a
a k i
µv800
0
-600
0
+600msec
/əpiˈkapə/
/əpaˈkipə/
KSH
i k a
a k i
µv500
0
-500
0
+500msec
/əˈpikapə/
/əˈpakipə/
FBB
i k a
a k i
µv700
0
-500
0
+500msec
/əˈpikapə/
/əˈpakipə/

FIGURE 1

Whenever the sequence involves moving from a more open to a less open vocal tract (i.e., /-ik-/), only one peak of activity is present. On the other hand, when the sequence involves moving from a less open to a more open vocal tract (i.e., /-ki-/) two separate peaks of activity are present. One explanation for this might be that the /k/ gesture is anticipated during the /i/ in the /-ik-/ sequence and the motor commands for the two gestures merge into one--that is, the further closing required for /k/ during /i/ is anticipated. The converse is not true: the /i/ gesture during the /k/ in the /-ki-/ sequence is not anticipated, since the /i/ articulation is contradictory to the more closed vocal tract of /k/. Both cases would seem to fit Henke's model, as it has been extended to the motor command level.

Some other data were also inspected in this light (see Figure 2). These data were again EMG recordings from the genioglossus muscle. They were recorded as the subject repeated a series of four-syllable nonsense words beginning and ending with schwa. The two medial vowels were /i/, and the stress was systematically varied between the first and second /i/. Bilabial consonants were used since the production of a bilabial consonant is not expected to interfere with the preservation of the lingual articulation for /i/. The first and second consonants were /p/ and the final consonant was systematically varied between /p/ and /b/, producing four utterance types that were then repeated at slow and fast rates. (Some of the design detail here is to allow analysis of these data for other purposes.) The data presented are from one speaker of American English, and roughly parallel data have been obtained from two other speakers.

The EMG recordings were inspected to determine whether the two sequences of vowels (the first: stressed-to-unstressed; the second: unstressed-to-stressed) that are separated by an intervening bilabial articulation were like the /-ki-/ and /-ik-/ sequences examined earlier. Although the intervening /p/ is presumably not contradictory to the maintenance of the /i/ vowel articulation, we see that there are two separate peaks of activity in every condition (Figure 2).

We examined the two stress conditions to see if the hypothesis advanced earlier, that EMG activity will merge for sequences moving to a more closed vocal tract, is supported. The vocal tract might be expected to be more closed for a stressed /i/ than for an unstressed /i/. Thus, we might expect to find less separation of the two peaks of EMG activity in the condition where the second vowel is stressed. In fact, the lowest valleys between vowel peaks occur for the unstressed-to-stressed sequences. The most obvious explanation for this difference is that the duration of the /p/ closure is longer before a stressed vowel than before an unstressed vowel and so the EMG signal falls to a lower level before the second vowel begins.

Our conclusion, then, is twofold: first, the /-ik-/ and /-ki-/ sequences are not part of the same subset of data as the /ipi/ utterances; second, Henke's model does not hold at the EMG level in an example where we might have expected it--for two vowels separated by a nonantagonistic consonant gesture. Features are not anticipated as soon as they are no longer contradictory to intervening segments.

In summary, Henke's look-ahead model of anticipatory coarticulation predicts that an articulatory feature will be anticipated as soon as it is no longer contradicted in the intervening speech string. Examination of two sets of EMG data has revealed support for the model at the motor command level, in one instance,

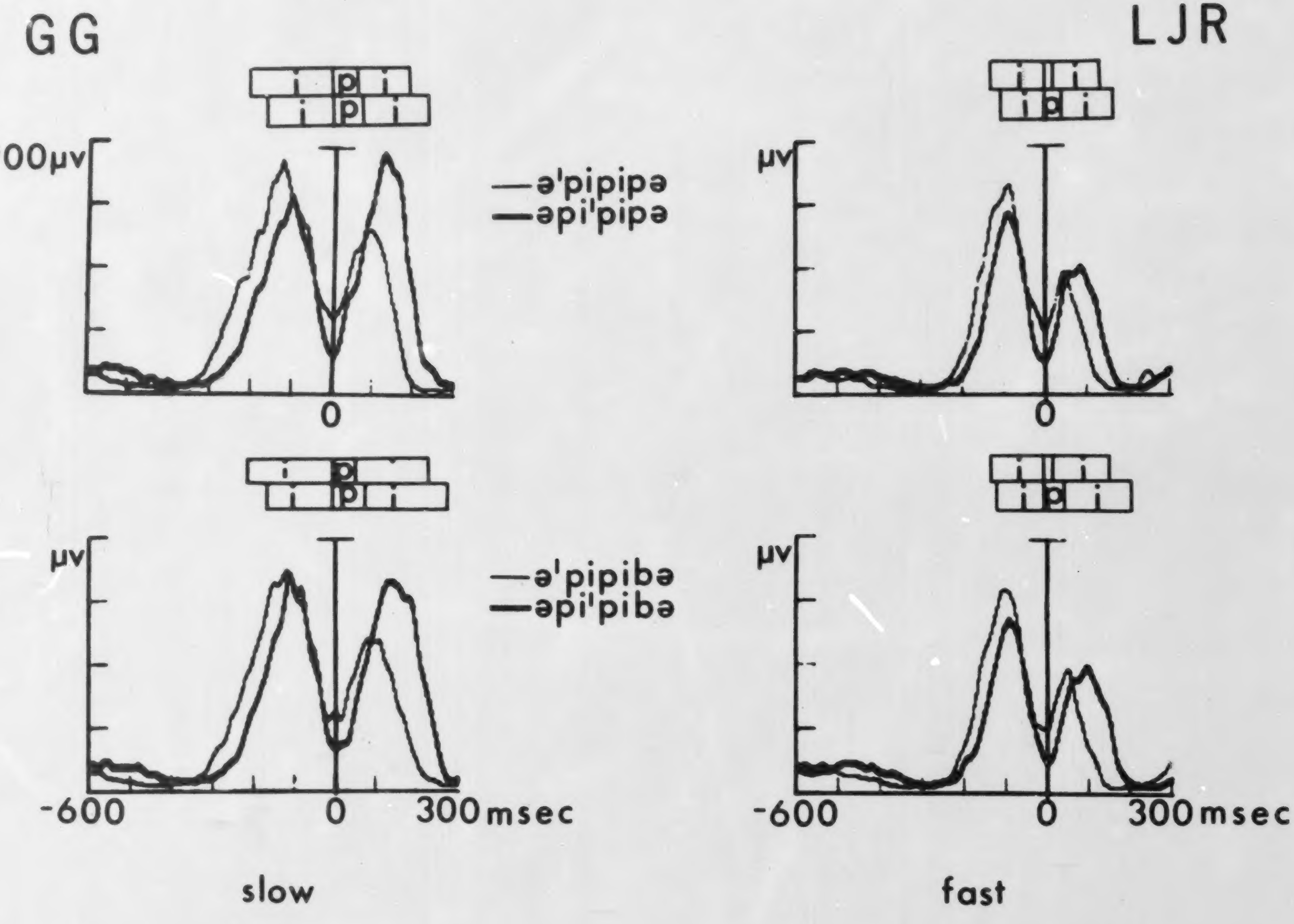

GG
LJR
900µv
µv
µv
µv
—ə'pipipə
—əpi'pipə
—ə'pipibə
—əpi'pibə
i p i
i p i
i p i
i p i
0
0
-600
0
300 msec
-600
0
300 msec
slow
fast

FIGURE 2

and a contradiction of the model in the other instance. In the former case, the interacting gestures were contiguous, while in the latter case, they were separated by a presumably noncontradictory consonant articulation.

REFERENCES

Bell-Berti, F. and K. S. Harris. (1973) The motor organization of some speech gestures. Haskins Laboratories Status Report on Speech Research SR-35/36, 1-5.

Henke, W. L. (1966) Dynamic articulatory model of speech production using computer simulation. Unpublished Ph.D. thesis, Massachusetts Institute of Technology.

Electromyographic Study of the Velum During Speech*

T. Ushijima[+] and H. Hirose[+]
Haskins Laboratories, New Haven, Conn.

INTRODUCTION

In a previous study, direct viewing of the velum by use of a fiberoptic system revealed several interesting findings on the velopharyngeal mechanism during speech articulation (Ushijima and Sawashima, 1972). Our interest was then directed to an investigation of the relationship between the actual movements of the velum and their motor commands during speech. For this purpose an electromyographic (EMG) study of velar movements was undertaken, using test words similar to those used in the earlier fiberoptic study. This procedure was intended to offer EMG data comparable with those of velar movements.

In this report we will discuss EMG activity of the levator palatini in relation to the results of the earlier fiberoptic experiment. The levator palatini has been generally considered to be the principle muscle of velopharyngeal closure. The aim in this study was, therefore, to investigate the possible correlation between levator activity and apparent velar height, especially for nasal coarticulation.

PROCEDURE

Two Japanese speakers (HH and TU), both of Tokyo dialect, served as subjects for this EMG experiment. They had not served as subjects in the earlier fiberoptic experiment. The subjects read a randomized list of 28 utterance types 16 times (Table 1). In this table a syllable-final nasal is indicated as /N/, while a syllable-initial nasal is shown as /n/.[1] The test words, consisting of meaningful disyllabic words, were included in a carrier sentence of /----desu/ (it is ----). None of the test words contain any accent kernel. The subjects were required to read the sentences at a conversational rate, which proved to be nearly identical for the two different experiments.

---

*Paper presented at the 87th meeting of the Acoustical Society of America, New York, 26 April 1974.

[+] On leave from University of Tokyo, Japan.

[1] In Japanese the syllable-final nasal /N/ is characterized by some special features. The phoneme /N/ has a duration equal to that of one mora. The specification of the articulation for this segment seems entirely dependent on that of the following phoneme.

TABLE 1: List of test words.

	For EMG (meaningful words)	For Filming (meaningful words)	(nonsense words)
1)	/see'ee/	/see'ee/	/aiueoaiueoa/
2)	/seesee/		/tetetete/
3)	/seetee/		/sesesese/
4)	/seezee/		/dededede/
5)	/teetee/		/zezezeze/
6)	/zeesee/		/nenenene/
7)	/see'eN/	/see'eN/	/teNteNteNteN/
8)	/seeseN/		
9)	/tee'eN/		
10)	/teeteN/		
11)	/teedeN/		
12)	/teezeN/		
13)	/teenee/		
14)	/deenee/	/deenee/	
15)	/seN'ee/	/seN'ee/	
16)	/seNsee/	/seNsee/	
17)	/teNtee/		
18)	/zeNsee/	/zeNsee/	
19)	/deNsee/		
20)	/neNsee/	/neNsee/	
21)	/seeneN/	/seeneN/	
22)	/teeneN/		
23)	/seN'eN/	/seN'eN/	
24)	/seNseN/		
25)	/teNteN/		
26)	/seNneN/	/seNneN/	
27)	/teNneN/		
28)	/neNneN/		
		/teNseN/	
		/deNseN/	
		/heNseN/	
		/seNteN/	
		/'eNseN/	
		/seNdeN/	
		/'eNsee/	

Conventional hooked-wire electrodes were inserted into the levator muscle perorally (Hirose, 1971). The EMG signals were computer-averaged with reference to a line-up point on the time axis. A more detailed description of the computer-processing system used is reported elsewhere (Kewley-Port, 1973a).

RESULTS AND DISCUSSION

1) Vowel and Nonnasal Consonant

Figure 1 shows three examples of averaged EMG curves for the two subjects. The thin line represents a /CVV'VV/ sequence, /see'ee/, with a syllable boundary occurring within the four successive vowel phonemes.[2] The thick line represents a /CVVCVV/ sequence, /seesee/. The dashed line represents a /CVVCVN/ sequence, /seeseN/, with a syllable-final nasal at the end of the second syllable. Zero on the time axis is the voice onset of /e/ after the initial /s/, which was obtained from the audio signal and which served as the line-up point for averaging. It is clear that the level of EMG activity for the vowel /e/ is much lower than that for /s/.

Kewley-Port (1973b) described the results of an experiment with multiple electrode insertions to different locations in the levator palatini muscle. She commented that consistent patterns of averaged EMG curves, with high correlations for each of several different electrode locations, were obtained regardless of different amplitudes of the maximum scale values among them. Therefore, the activity pattern picked up from one location should represent the overall change in the motor command to this particular muscle. In this sense, the different levels of activity between /s/ and /e/, shown in Figure 1, lead us to assume that there are quantitatively different neural commands for movements of the velum for consonant and vowel production.

Even if nasality may be considered as a "one muscle-one parameter" system, a decrease in levator activity for the vowel /e/ should not necessarily be interpreted as indicating a proportional decrease in absolute velar height. Instead, the amount of EMG activity is known to be proportional to the mechanical work required to approximate the required articulatory configuration (MacNeilage, 1972).

Bell-Berti and Hirose (1972a) pointed out a similar moderate difference in EMG potentials, between /i/ and /b/, and stated that such a difference may not be sufficient to result in a considerable shift in velar height, while the far greater increase in EMG activity for an oral consonant following a nasal will be sufficient to cause a considerable shift in velar height. In our earlier fiberoptic data, all the /see'ee/ samples [an example is shown in Figure 2 (a)] indicate that the velum stays at an almost constant height throughout the test word, or shows only a very slight decrease corresponding to the vowel portion of the test word. Thus, the activity level for /e/ in Figure 1 seems sufficient to maintain the velar height after having once reached the height for /s/.

---

[2] According to Hattori (1961), "'" represents a sort of consonant phoneme which has no manifest articulatory characterization except that it may indicate a syllable boundary.

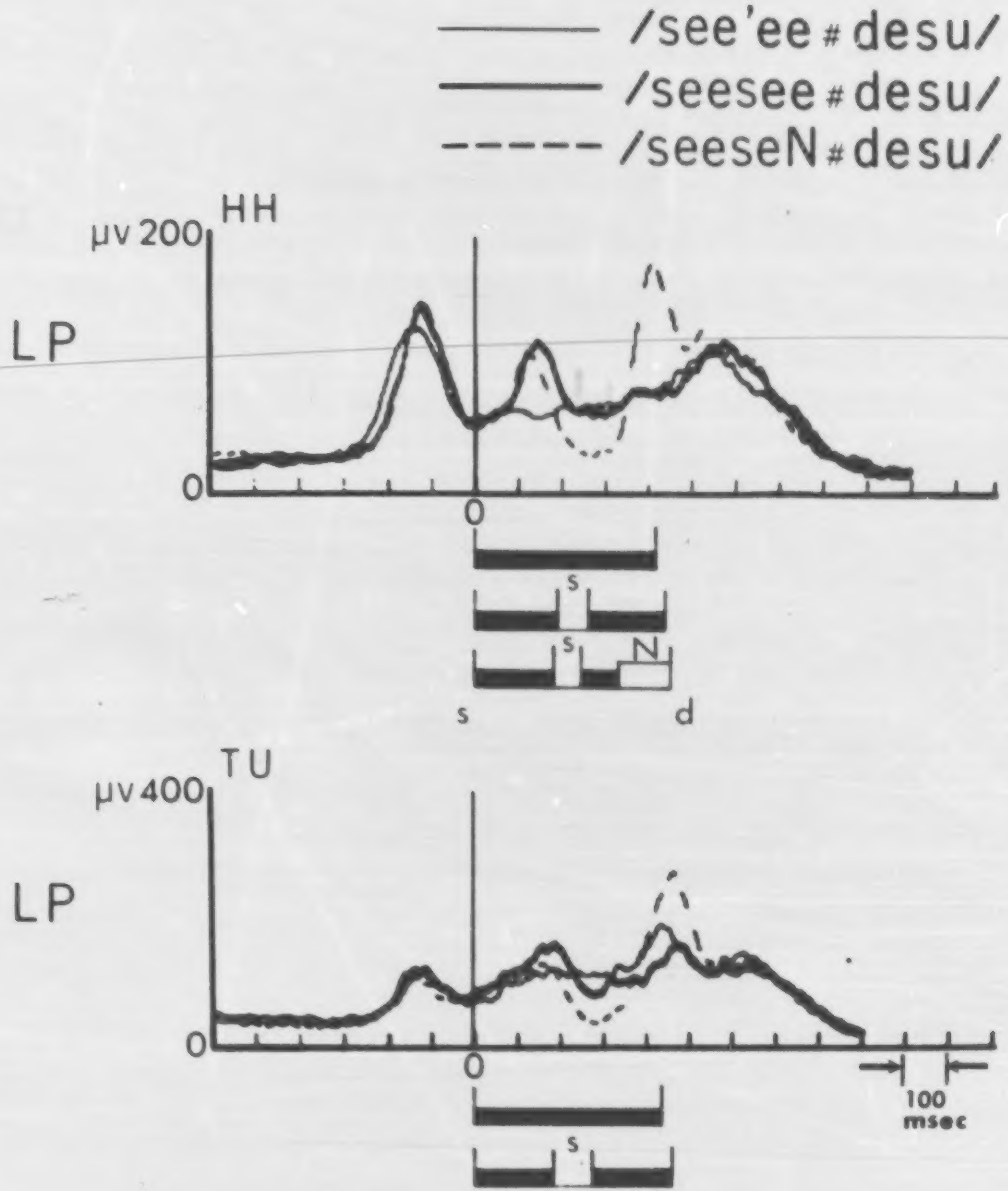


Figure 1: Superimposed averaged EMG curves for three utterance types: /see'ee#desu/ (thin line), /seesee#desu/ (thick line), and /seeseN#desu/ (dashed line). "'", "#", and "N" represent, respectively, a syllable boundary, a word boundary, and a syllable-final nasal.

FIGURE 2

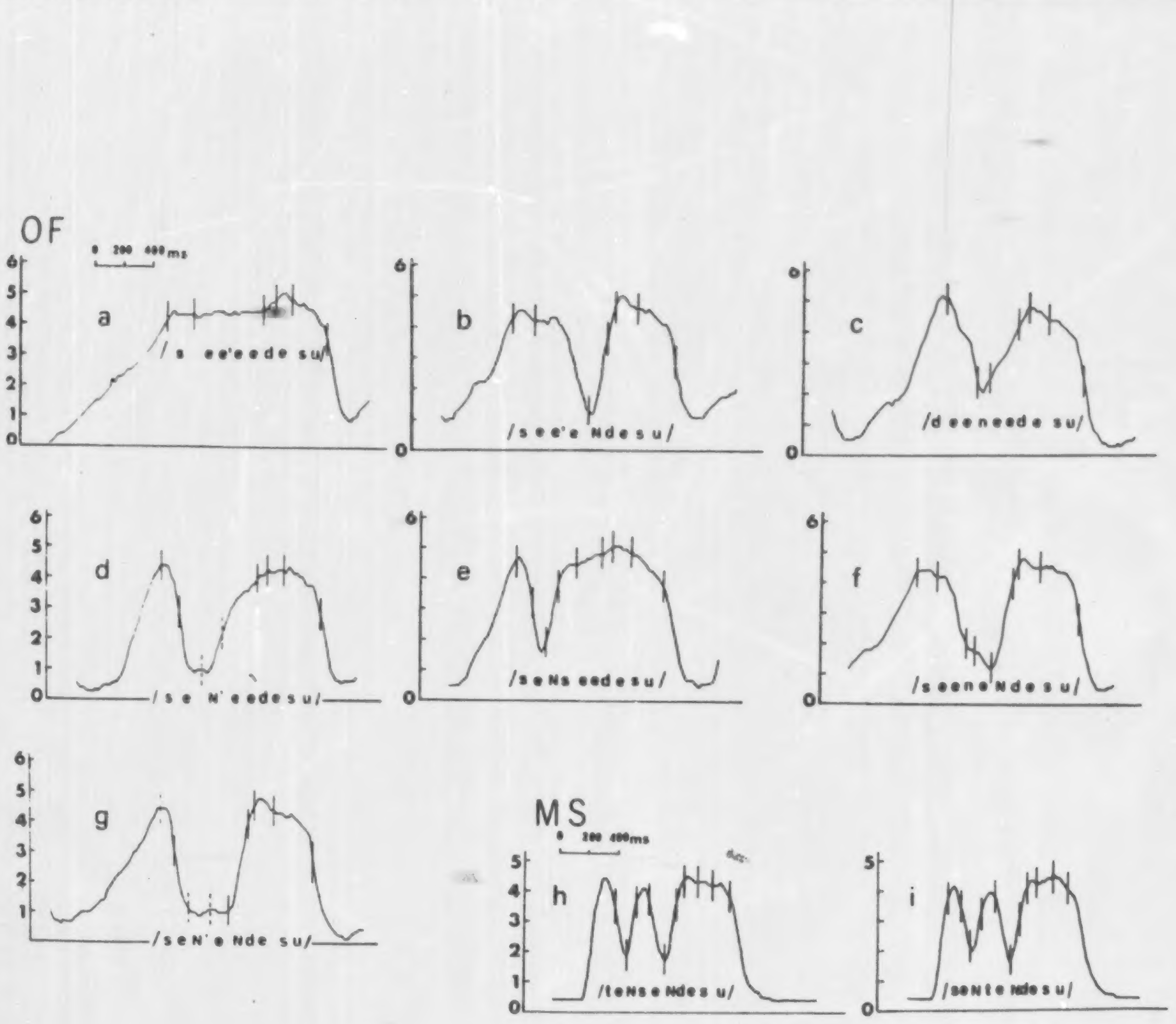


Figure 2: Selected examples of the velum height obtained from the previous fiberoptic study: (a) /see'ee/, (b) /see'eN/, (c) /deenee/, (d) /seN'ee/, (e) /seNsee/, (f) /seeneN/, and (g) /seN'eN/ from Subject OF; (h) /teNseN/ and (i) /seNteN/ from Subject MS. The ordinate is an arbitrary linear scale. The subjects (OF and MS) were different from those who served in the present EMG study.

In Figure 1, Subject HH shows a higher peak value for the initial /s/ than for the intervocalic /s/, while Subject TU does otherwise.[3] This fact would suggest that the neural commands necessary for the velopharyngeal closure for the two /s/s appear to differ in degree between the two subjects.

If we assume that velar height is generally constant during the repetition of CV syllables, the difference in muscle activity seen between word-initial and intervocalic /s/ in Figure 1 is considered not to be transformed into a clear difference in absolute velar height. In other words, subtle variance in the neural input to the velum is more directly reflected in the time course of the averaged EMG activity than in the time course of the actual velar movement. The averaged EMG may be substantially influenced by other factors. Some suprasegmental factors such as the existence of an accent kernel or stress, for example, may well affect the EMG level, even though such factors might not be completely realized in velar height.

For /N/, levator activity falls to a level observed for the resting state.

We may infer from our EMG data that the neural signal to the velum is not controlled by a simple dichotomy, such as an on-off mechanism. Instead, the absolute activity level for a given nonnasal phoneme may vary with the phonetic environment.

2) Differences Among Four Nonnasal Consonants

Figure 3 compares averaged levator EMG activity for the consonants /t/, /s/, /d/, and /z/. Although the material does not cover all possible combinations of the four consonants, the two nasal consonants, and the vowel /e/, we can evaluate the peak values for each oral consonant in comparable phonetic environments. The consonant pairs are selected as follows:

1) /teenee/ vs /deenee/
/teeteN/ vs /teedeN/ for /t/-/d/ comparison

2) /seesee/ vs /zeesee/
/seNsee/ vs /zeNsee/
/seesee/ vs /seezee/ for /s/-/z/ comparison

3) /tee'eN/ vs /see'eN/
/teeneN/ vs /seeneN/
/teNneN/ vs /seNneN/
/seetee/ vs /seesee/ for /t/-/s/ comparison

4) /deNsee/ vs /zeNsee/
/teedeN/ vs /teezeN/ for /d/-/z/ comparison

---

[3] Immediately preceding each utterence the subjects were required to inspire through the nose. Therefore, the flat portion of the averaged EMG curves before the peak for the initial /s/ is considered to correspond to the resting state of the velum. Thus, both subjects do not seem to have any "speech ready" position for the velum, i.e., the velum appears to move smoothly from the resting position to the position for the initial velopharyngeal closure.

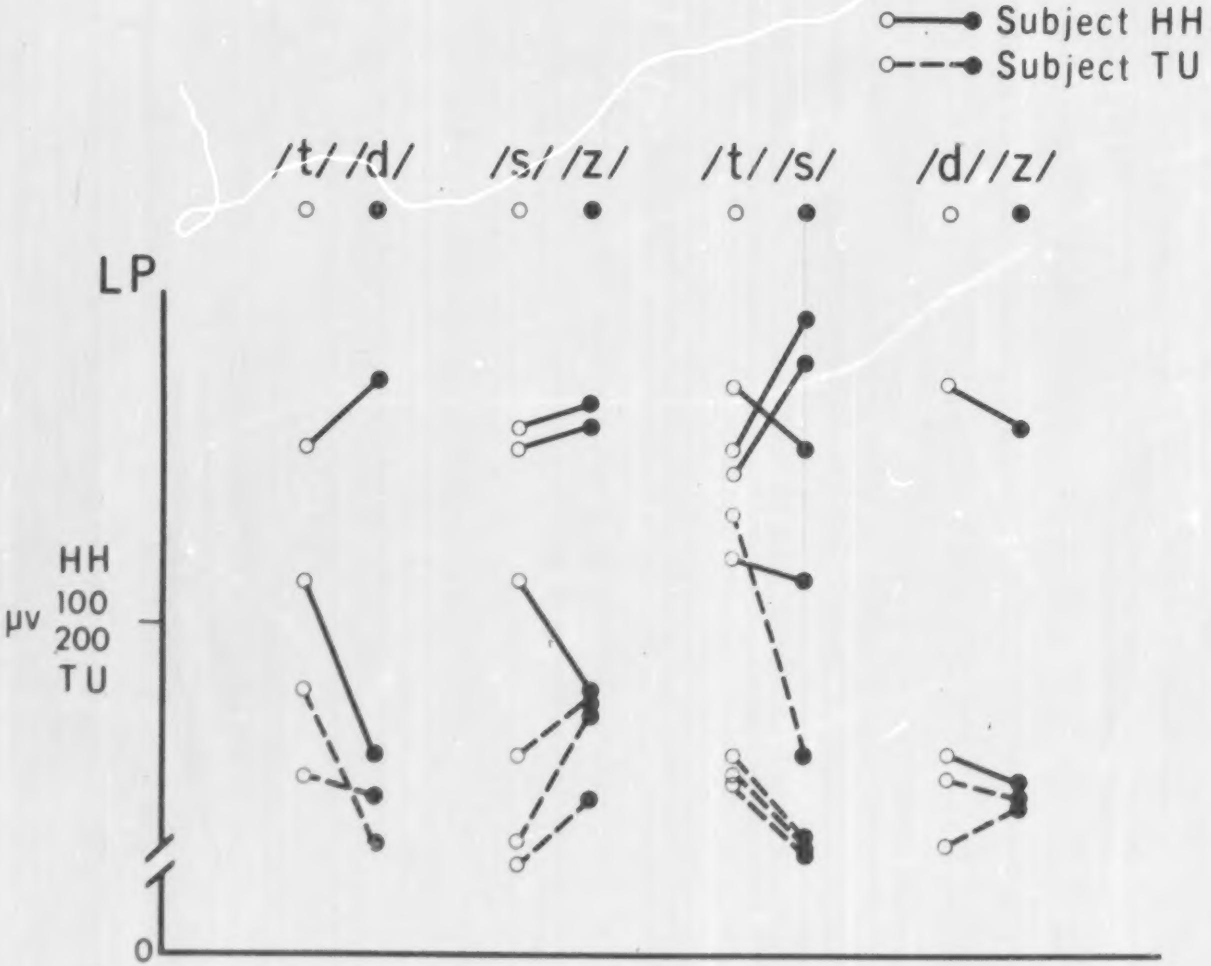


FIGURE 3

Figure 3: Comparison of levator activity for nonnasal consonant pairs (1).

In the figure, the consonant pairs are connected with solid lines for Subject HH and with dotted lines for Subject TU, and all the comparisons are shown pairwise regardless of the difference in the position of the consonants in the test words.

It is obvious that even in the same subject there is no consistent difference in peak activity between either voiced-voiceless or stop-fricative pairs. This holds true even when the consonant position is taken into consideration, though the detail of that is not shown in the figure.

Bell-Berti and Hirose (1972b) found differences in levator activity associated with stop consonant voicing for two out of three subjects. For their subjects, the presence or absence of differences in levator activity could be explained on the basis of intersubject differences in strategy for velopharyngeal enlargement to maintain voicing. Since data were obtained only from the levator in the present study, it is impossible to decide whether the two Japanese subjects use a cavity enlargement strategy that does not involve the levator, or alternatively, whether there is a difference between Japanese and English. The lack of difference between peak height for stops and fricatives was also observed in the earlier fiberoptic study.

The wide variation in peak values in Figure 3 may be due to the effect of a difference in phonetic environment. The effect is further investigated in Figure 4, where utterances are classified into seven groups according to the contextual construction, indicated as /C_1ee---/, /C_2eN---/, /C_3een--/, /--eC_4ee/, /--eC_5eN/, /--NC_6ee/, and /--NC_7eN/. Open circles (Subject HH) and filled circles (Subject TU) indicate the peak values for the /C/s. The mean of those values within each group is shown by a short horizontal line (the solid line for Subject HH and dotted line for Subject TU).

For both subjects, consonants in absolute initial position show the same peak height, whether there is a following nasal consonant of either type. The subjects differ in that for TU consonant position change has no effect, while for HH it does have an effect. Although such individual differences in EMG peak value for consonants may not be directly related to the difference in velar height, it is interesting to note that the two subjects show the rather different patterns described above.

The second finding worthy of note is that the peak values for C_6 and C_7 are far greater than those for the other groups. This would suggest that the neural command is organized so that the muscle activity is greatly increased for elevating the velum immediately after it has been lowered for the preceding /N/ segment. Bell-Berti and Hirose (1972a) asserted that there is a strong correlation between the magnitude of the increase in EMG potential and the magnitude of the change in velar height. The increased EMG potential after /N/ in these cases may be reasonably explained as being essential for the longest excursion of the velum from the position near the resting state to the elevated position for the succeeding stop consonants.

In this respect, our earlier film analysis of the velum indicates that velar height for the consonants after /N/ is no greater than that for the word-initial consonants in the test words [Figure 2 (e, h, and i)]. In light of this finding, then, the increased EMG for C_6 and C_7 is considered to indicate neither greater maximum elevation of the velum nor, presumably, tighter velopharyngeal closure,

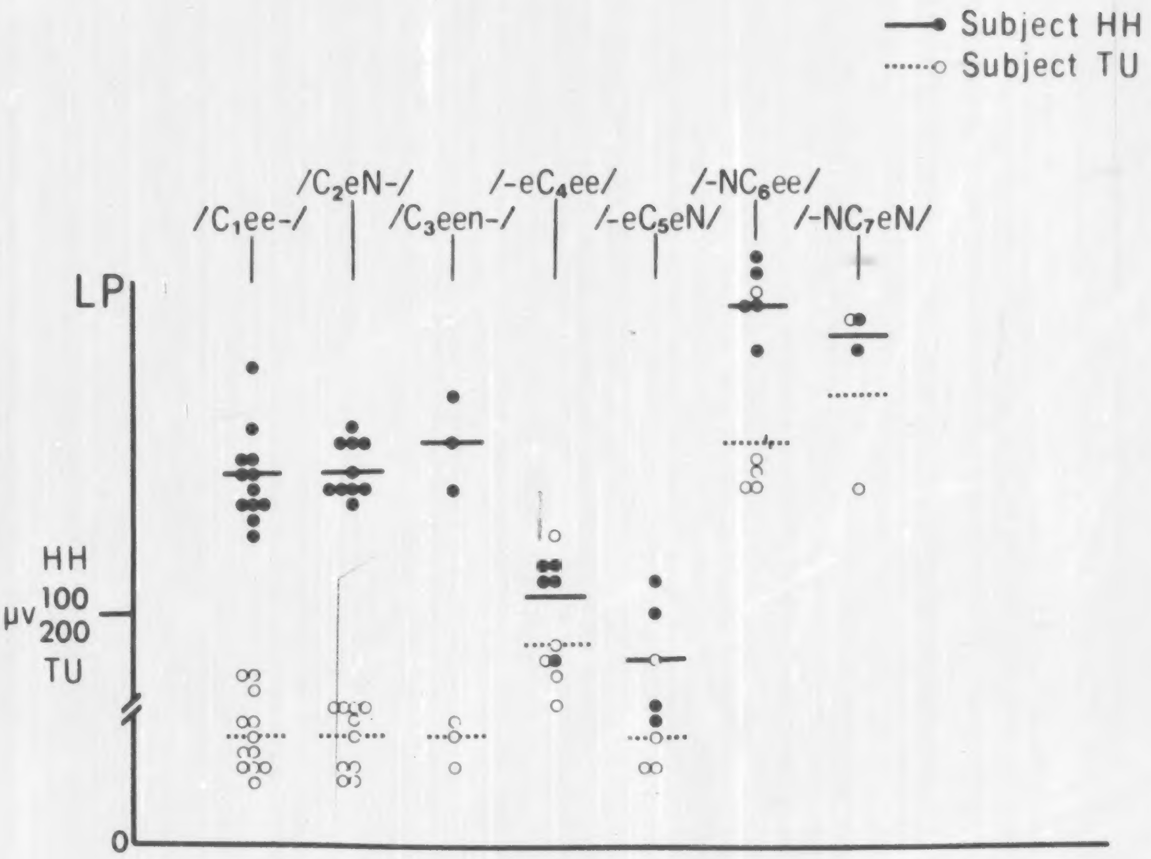


FIGURE 4

Figure 4: Comparison of levator activity for nonnasal consonant pairs (2).

but instead to indicate the contraction strength necessary to achieve adequate velopharyngeal closure for the oral consonants following nasals.

3) Difference Between /N/ and /n/

The earlier data, from velar movement analysis, implying an inherent difference between /n/ and /N/, with greater nasalization for the latter in Japanese (Figures 2 and 5), were compared with the EMG results obtained in the present study.

Figure 6 shows superimposed EMG curves for /teenee/ and /seN'ee/, each containing one nasal segment in intervocalic position. The downward slope of the EMG curves after the peak for the word-initial consonant is apparently steeper for /N/ than for /n/. This may be regarded as indicating greater speed of velar lowering for /N/ than for /n/. However, the apparently greater slope for /N/ might also be explained by the fact that the duration of the prenasal vowel segment is shorter before /N/. In any event, the minimum EMG activity preceding the nasal sounds is slightly lower for /N/ than for /n/ for both subjects. The question of whether or not the slight difference in activity level between /N/ and /n/ indicates a difference in actual velar height can be answered by combining EMG recordings with fiberoptic observation on the same subject, a process that is now in progress.

The segmental duration of /N/ is clearly longer than that of /n/ (Figure 6). The nasal segment duration may have some relationship to the observed differences in velar height. Ohala (1971) stated that the palate lowers more for word-final nasal consonants than for word-initial nasals, but he did not comment on the difference in the duration of nasal segments. Further studies, using different speaking rates and measurements of nasal segment duration in various phonetic positions, seem to be needed. We also see in Figure 6 that levator activity remains suppressed after /n/ but not after /N/, a fact that will be further discussed below.

4) Coarticulatory Movements of the Velum

Many authors have described the coarticulation of nasality, based on cineradiographic observations of velar movement, on aerodynamic studies, and on acoustic analyses of speech. If coarticulation is interpreted as "the influence of one speech segment upon another speech segment" (Daniloff and Hammarberg, 1973), we might expect to observe such effects at the level of the motor command, or electromyographically. Dixit and MacNeilage's (1972) EMG and aerodynamic studies of Hindi on the extent of coarticulatory effects are so unique as to lead to the following conclusions:

1) The effect of coarticulation can stretch across four segments.

2) Carry-over effects (left-to-right effects) are as extensive as anticipatory effects (right-to-left effects).

3) The temporal scope of coarticulatory effects is unrestricted by syllable or word boundaries.

Our data on Japanese are not entirely consistent with their results.

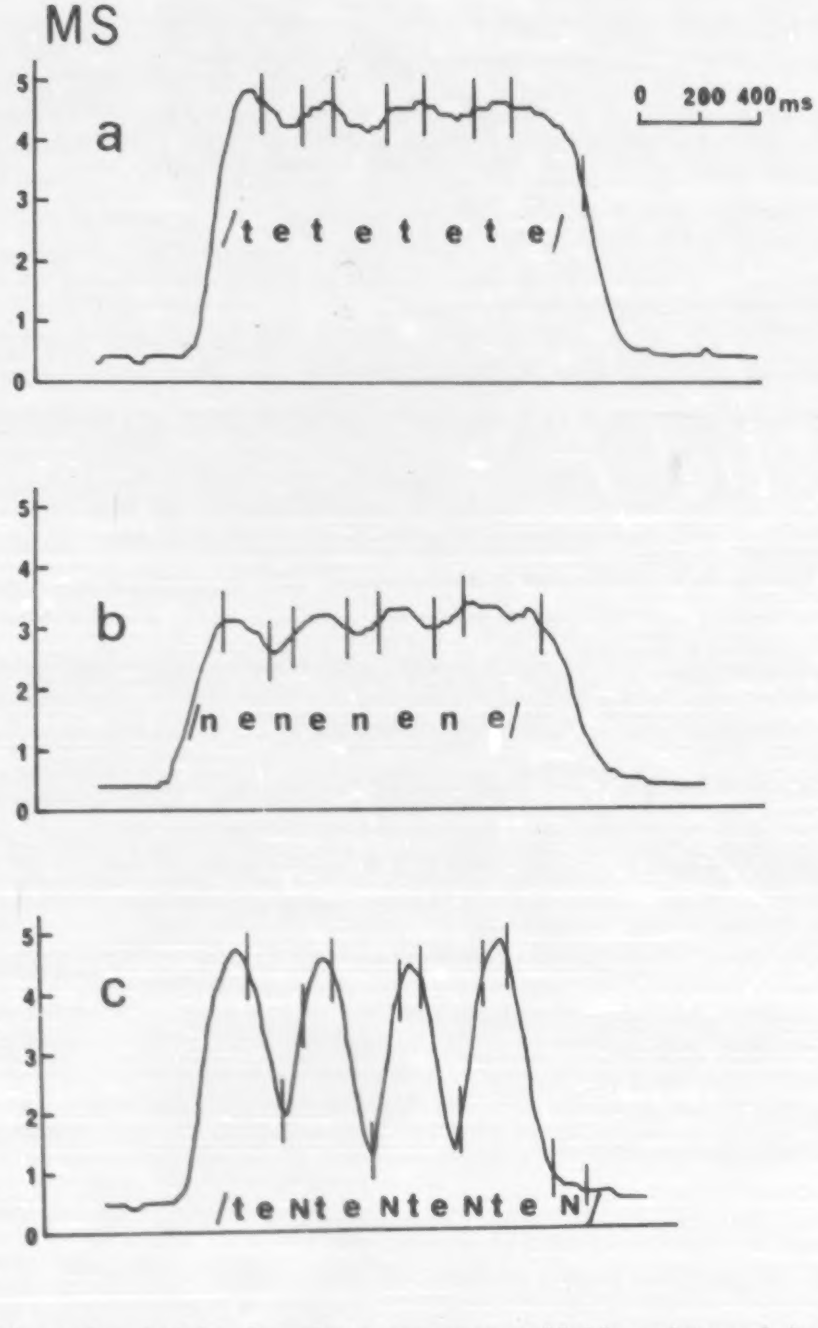


Figure 5: The velum height for three nonsense syllables observed in Subject MS (cf. Figure 2): (a) /tetetete/, (b) /nenenene/, and (c) /teNteNteNteN/. The ordinate is an arbitrary linear scale.

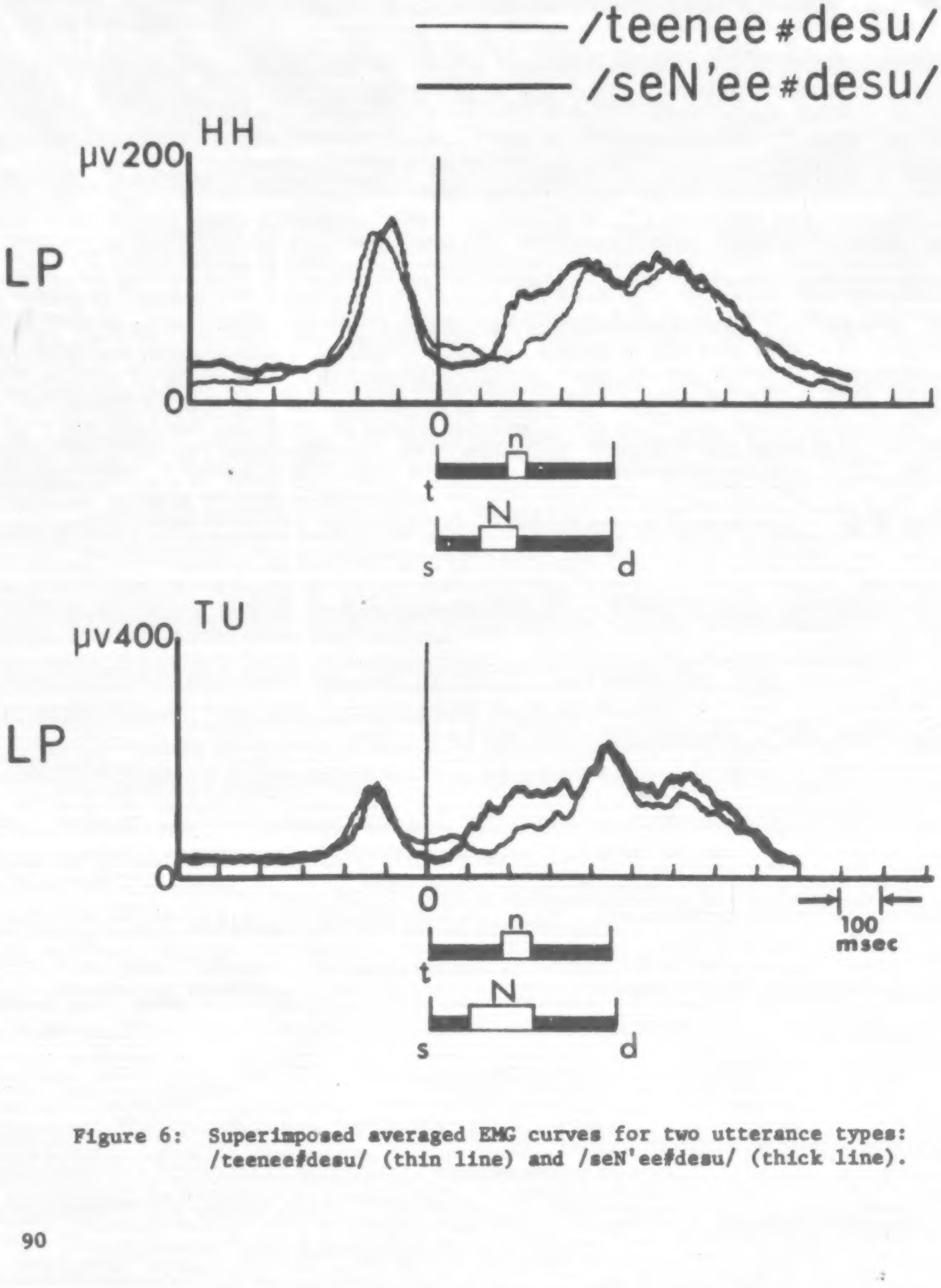


Figure 6: Superimposed averaged EMG curves for two utterance types: /teenee#desu/ (thin line) and /seN'ee#desu/ (thick line).

A) Carry-over coarticulation. EMG activity for the underlined vowel segments of the second syllable in /seN'ee/ (Figure 7) does not reveal any carry-over suppression from the preceding /N/. Rather, it shows a far greater increase than the EMG level necessary for the underlined vowel sounds in /see'ee/. If the carry-over effect represents a change in the time course of the neural command, as indicated by Dixit and MacNeilage's (1972) EMG data on Hindi, we should expect decreased activity for the underlined segments in /seN'ee/. But this is not the case. Phonemically there is no contrastive nasality in Japanese vowels, so presumably, there are no restrictive influences against velar lowering. At the level of the neural command to the velum the carry-over effect, if it is observed at all, is not realized as decreased muscle activity. In this case, carry-over coarticulation does not seem to extend beyond the syllable boundary between the two syllables of the test words.

On the other hand, comparison of the /see'ee/-/teenee/ pair in Figure 8 shows a clear carry-over effect of a syllable-initial /n/ on the following vowel segments in the second syllable. Specifically, the activity for the /ee/ after /n/ in /teenee/ never surpasses the level for /'ee/ in /see'ee/. This is also evident in the difference in activity level for the post-/n/ vowel segments shown in Figure 6. A possible explanation for this difference is that the vowel segments after a syllable-final /N/ may have to be oralized to prevent listener confusion. Further comment will be made below on the restrictions on coarticulation.

Although the carry-over effect does not appear to be present in the case of a /CeN'ee/ sequence at the motor command level, the earlier fiberoptic study showed a lower velar height for the vowel segments following /N/ than for vowel segments in oral environments. It should be reasonable to assume, therefore, that realization of the carry-over effect in the form of velar movement in those cases may be due to some inherent mechanical response characteristics of the velum. At present, we would agree with the speculation (Daniloff and Hammarberg, 1973) that carry-over coarticulation is partly due to mechano-inertial limitations on the articulators as a physical system.

B) Anticipatory coarticulation. EMG evidence supports the existence of anticipatory nasal coarticulation in vowel production. Figure 7 compares three utterance types, /see'ee/, /seN'ee/, and /seeneN/. Unless EMG activity for /e/ after /s/ is influenced by anticipatory effects from /N/ in /seN'ee/ or from /n/ in /seeneN/, the three curves should show the same level of activity, at least for the short period following the peak for the initial /s/. In this respect, however, there is a clear difference among the examples in Figure 7, where the curve for /see'ee/ shows a higher EMG level before the line-up than the other curves for Subject HH. Subject TU shows a similar tendency between /see'ee/ and /seN'ee/.

Another example indicating the anticipatory effect in the vowel segment before /N/ is shown in the /seesee/-/seeseN/ comparison in Figure 1. The effect appears to manifest itself as about a 25 msec difference in timing of initiation of EMG suppression after the peak for the intervocalic /s/, which is significant even when the difference in timing of postconsonantal vowel onset is taken into consideration. It seems reasonable, then, to conclude that the neural commands for vowels followed by nasals are reorganized by the anticipatory effect.

As far as we have surveyed the collected data, the anticipatory effect and the carry-over effect have different characteristics at the EMG level. The

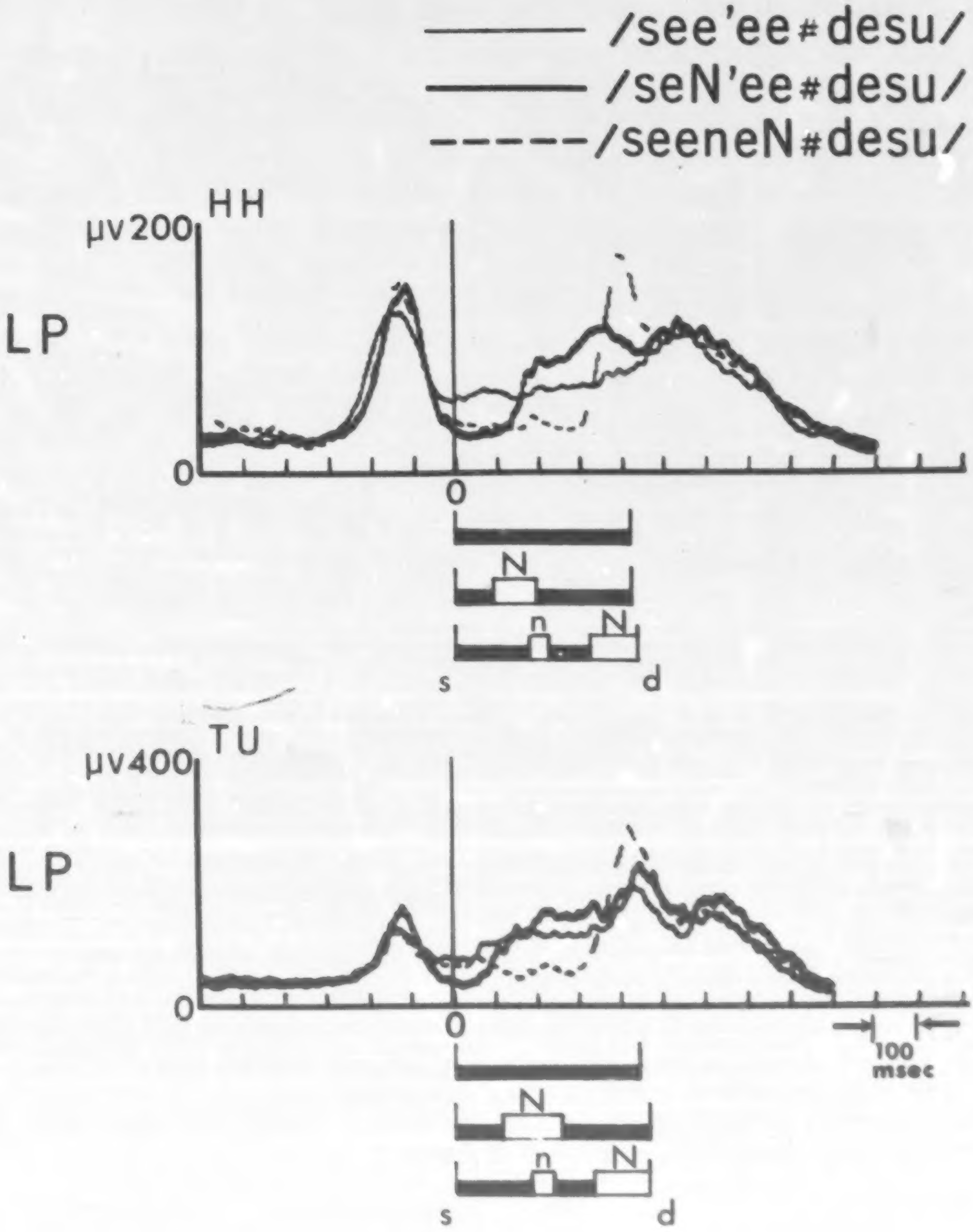


Figure 7: Superimposed averaged EMG curves for three utterance types: /see'ee#desu/ (thin line), /seN'ee#desu/ (thick line), and /seeneN#desu/ (dashed line).

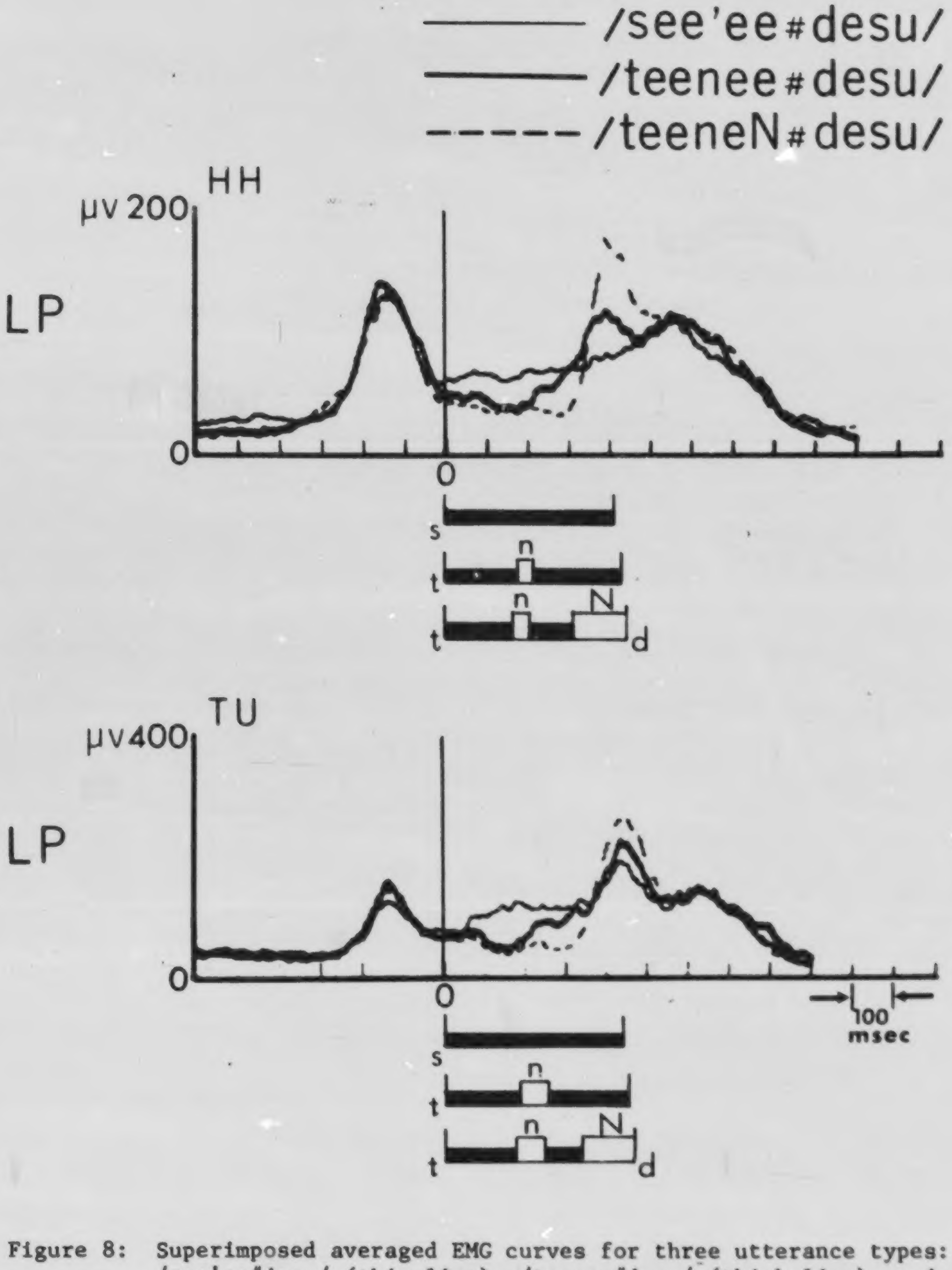


Figure 8: Superimposed averaged EMG curves for three utterance types: /see'ee#desu/ (thin line), /teenee#desu/ (thick line), and /teeneN#desu/ (dashed line).

anticipatory effect manifests itself as some kind of reorganization of the neural command following the so-called "look ahead" principle or future scanning mechanism, to which we will refer later. The carry-over effect manifests itself as some kind of reorganization of the time course of the neural command for vowels following some nasal segments. After syllable-final /N/, there is no carry-over effect at the EMG level. In this sense, the carry-over effect is less pervasive than the anticipatory effect. In such cases, realization of the carry-over effect as actual velar lowering is due, in part, to some mechanical response characteristics of the velum.

C) Restriction of coarticulation. In the examples such as /teenee/, /deenee/, /seeneN/, and /teeneN/, the syllable boundary before /n/ does not restrict the anticipatory coarticulation from /n/ in the second syllable. Figure 8 compares the /teenee/-/teeneN/ pair with /see'ee/. In the figure, after the peak for the initial consonant, activity for the underlined /ee/ in each case with a syllable-initial nasal is somewhat more suppressed than in the case of /see'ee/. The higher degree of suppression or greater decrease in EMG level begins as early as about 50 msec before the line-up point for Subject HH, and about 50 msec after the line-up point for Subject TU. Many authors now agree with the opinion that anticipatory coarticulation of nasality extends across the syllable boundary, and our present results are partly consistent with that opinion.

The next important finding in this study is evidence of a restriction of the anticipatory effect of velar lowering. Moll and Daniloff (1971) have suggested, following Henke (1966, quoted in Moll and Daniloff, 1971), that anticipatory coarticulation operates on a "look ahead" principle. A feature is articulated in a speech string as soon as it can be. Thus, if the vowels are presumed to be neutral with respect to nasalization, velar lowering for a terminal nasal should occur at the beginning of a preceding vowel string. The length of the vowel string should be irrelevant.

This hypothesis was tested by comparing three speech strings. Figure 9 shows three examples of averaged EMG curves for the two subjects. The thin line represents a /CVV'VV/ sequence, /see'ee/. The thick line represents a /CVV'VN/ sequence, /see'eN/, with a syllable-final nasal at the end of the second syllable. The dashed line represents a /CVN'VV/ sequence, /seN'ee/, with an /N/ at the end of the first syllable. About 150 msec before the line-up, there is always a peak for the high velum consonant /s/. Immediately after the peak, there is suppression of EMG activity in /seN'ee/ (the dashed line), indicating, of course, decreased activity for the syllable-final nasal. By contrast, in /see'eN/ (the thick line) the activity for the initial vowel segment after /s/ has the same level as the vowel in the utterance without the nasal. The activity begins to fall about 100 to 150 msec after the line-up.

If Moll and Daniloff's hypothesis of "unspecified" velar position for the vowel is applicable to Japanese vowels, the EMG signal for the vowel segment after /s/ in /see'eN/ should show the same decrease as for the underlined /e/ in /seN'ee/. However, the present data suggest that there is a restriction on anticipatory velar lowering.

In summarizing the results obtained so far from both direct viewing and EMG of the velum, there seems to be no anticipatory lowering of the velum during the first portion of the vowel segment in the /CVV'VN/ environment containing a syllable boundary (Figures 2 and 9). The results do not support Moll and Daniloff's

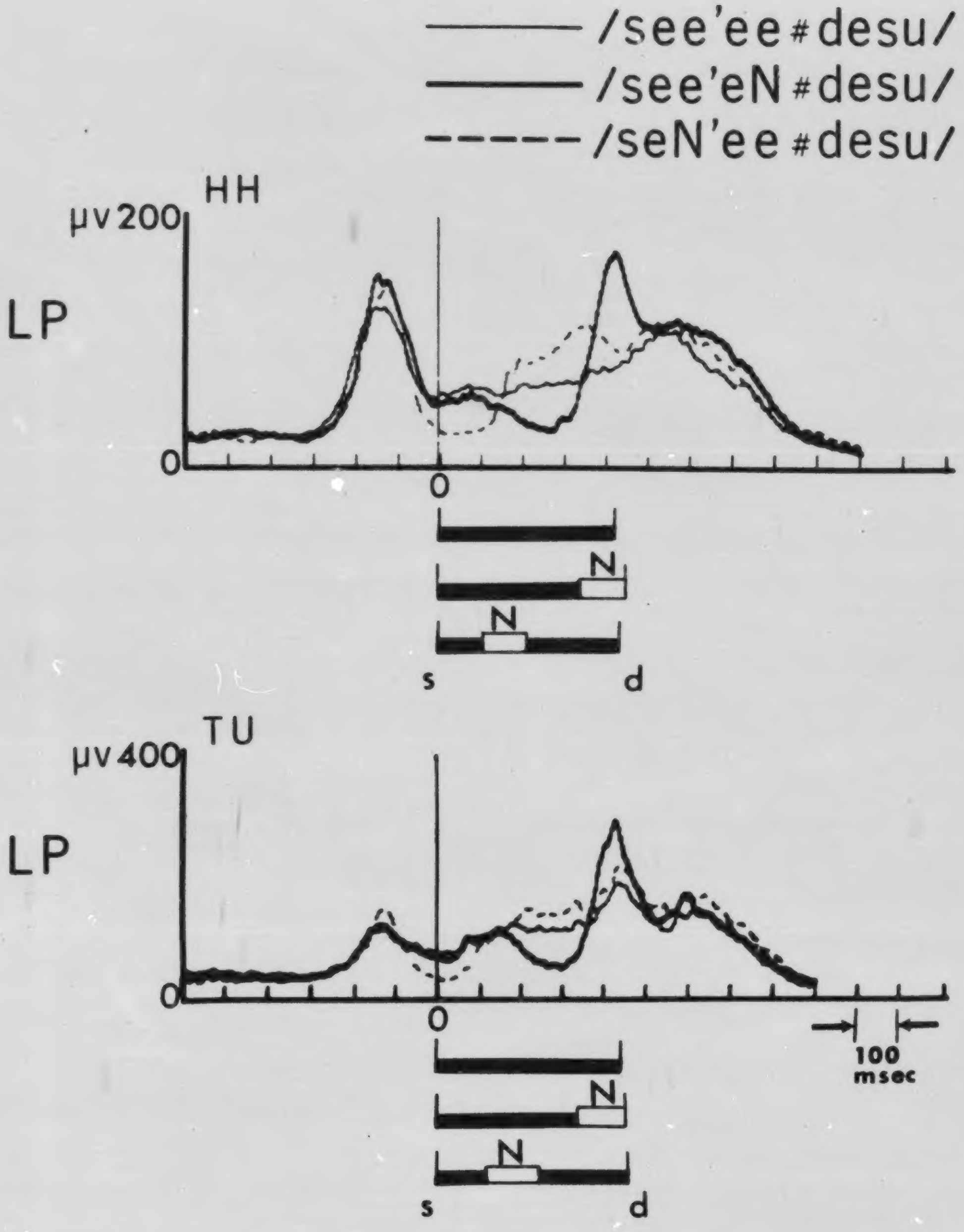


Figure 9: Superimposed averaged EMG curves for three utterance types: /see'ee#desu/ (thin line), /see'eN#desu/ (thick line), and /seN'ee#desu/ (dashed line).

(1971) proposal, although, to be sure, the languages tested are different. Since our disyllabic test words do not contain any obvious acoustic pause, we cannot explain this delayed onset of velar lowering by the existence of a prosodically marked grammatical boundary. This delayed onset of coarticulation might be due to a high-level reorganization of the input commands to the velum (McClean, 1973). Then, the presence of a syllable boundary within the vowel string and/or the number of the interposed vowel segments may well have some effect on anticipatory velar lowering.

Another possible explanation of this phenomenon might be the following. The specification of velar position for the elongation of Japanese vowels may not be neutral. Instead, the elongation of the vowel can be regarded as positively specified in terms of denasalization. Thus, coarticulation may not occur beyond the boundary, as in /see'eN/.

SUMMARY AND CONCLUSION

EMG recordings from the levator palatini muscle of two Japanese subjects lead us to summarize as follows:

1) From the viewpoint of the motor command level the velum is not controlled by a simple dichotomy such as an on-off mechanism.

2) There is no systematic segmental difference between either voiced and voiceless or stop and fricative consonants. However, the absolute activity level for a given nonnasal phoneme may not be predicted, but varies according to its context.

3) The different degree of nasalization for /n/ and /N/ seems to be realized electromyographically in the form of greater suppression of EMG activity for /N/.

4) There are different mechanisms for anticipatory and carry-over effects of coarticulation at the level of the motor command. The anticipatory effect is some kind of reorganization of the neural commands following the "look ahead" principle. On the other hand, the carry-over effect is less pervasive than the other in the sense that for the vowel segment after a syllable-final nasal there is no carry-over suppression of EMG activity.

5) There seems to be no anticipatory lowering of the velum during the vowel segments before a syllable boundary in the /CVV'VN/ environment. This phenomenon, which does not appear to support Moll and Daniloff's (1971) proposal, suggests that the elongation of vowels in Japanese might be regarded as positively specified in terms of denasalization.

REFERENCES

Bell-Berti, F. and H. Hirose. (1972a) Velar activity in voicing distinctions: A simultaneous fiberoptic and electromyographic study. Haskins Laboratories Status Report on Speech Research SR-31/32, 223-230.

Bell-Berti, F. and H. Hirose. (1972b) Stop consonant voicing and pharyngeal cavity size. Haskins Laboratories Status Report on Speech Research SR-31/32, 207-211.

Daniloff, R. G. and R. E. Hammarberg. (1973) On defining coarticulation. J. Phonetics 1, 239-248.

Dixit, R. P. and P. F. MacNeilage. (1972) Coarticulation of nasality: Evidence from Hindi. Paper presented at 83rd meeting of the Acoustical Society of America, Buffalo, N. Y.

Hattori, S. (1961) Prosodeme, syllable structure, and laryngeal phonemes. Studies in Descriptive and Applied Linguistics, Bulletin of the Summer Institute in Linguistics (International Christian University, Tokyo) 1, 1-27.

Hirose, H. (1971) Electromyography of the articulatory muscles; current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Kewley-Port, D. (1973a) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-183.

Kewley-Port, D. (1973b) Personal communication.

MacNeilage, P. F. (1972) Speech physiology. In Speech and Cortical Functioning, ed. by John H. Gilbert. (New York: Academic Press) 1-72.

McClean, M. (1973) Forward coarticulation of velar movement at marked junctual boundaries. J. Speech Hearing Res. 16, 286-296.

Moll, K. L. and R. G. Daniloff. (1971) Investigation of the timing of the velar movements during speech. J. Acoust. Soc. Amer. 50, 678-684.

Ohala, J. J. (1971) Monitoring soft palate movements in speech. Project on Linguistic Analysis (Department of Linguistics-Phonology Laboratory, University of California, Berkeley) 2, 13-27.

Ushijima, T. and M. Sawashima. (1972) Fiberscopic observation of velar movements during speech. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) 6, 25-38.

The Function of the Posterior Cricoarytenoid in Speech Articulation*

Hajime Hirose[+] and Tatsujiro Ushijima[+]
Haskins Laboratories, New Haven, Conn.

Participation of the posterior cricoarytenoid (PCA) muscle in laryngeal articulatory adjustments has been demonstrated by our previous electromyography (EMG) studies in which we observed increasing PCA activity for the production of voiceless segments in different languages (Hirose, 1971; Hirose and Gay, 1972; Hirose, 1973; Hirose, Lisker, and Abramson, 1973).

The aim of the present study is to investigate further the relationship between the pattern of PCA activity and glottal gestures for voiceless sounds of Japanese. Two separate experiments were performed: EMG of PCA, and fiberoptic observation of the glottis for the same sounds in the same subject.

A native Japanese subject of Tokyo dialect read randomized lists of meaningful Japanese words embedded in a frame sentence "soreo _____ to yuu" ("that we call _____"). Table 1 shows a list of the test words. These test words contain the voiceless fricative /s/ and stops /t/ and /k/ in word-initial and/or word-medial positions. They also contain devoiced vowel segments and voiceless geminate consonants. In this table, Q stands for geminates. The high vowel /i/ between two voiceless consonants is always devoiced.

TABLE 1: A list of test words used in the present experiment.

/seesee/	/kisee/	/seQsee/
/teetee/	/kitee/	/seQtee/
/keekee/	/kikee/	/seQkee/
/seetee/	/ki'ee/	/sekisee/
/seekee/		/sekitee/
/seki'ee/		/sekikee/

In the first part of the experiment, hooked-wire electrodes were inserted perorally into the PCA and into the interarytenoid (INT) as shown in Figure 1. The EMG signals were processed using the system described by Kewley-Port (1973).

---

*Presented at the 87th meeting of the Acoustical Society of America, New York, N. Y., April 1974.

[+] Also Faculty of Medicine, University of Tokyo, Japan.

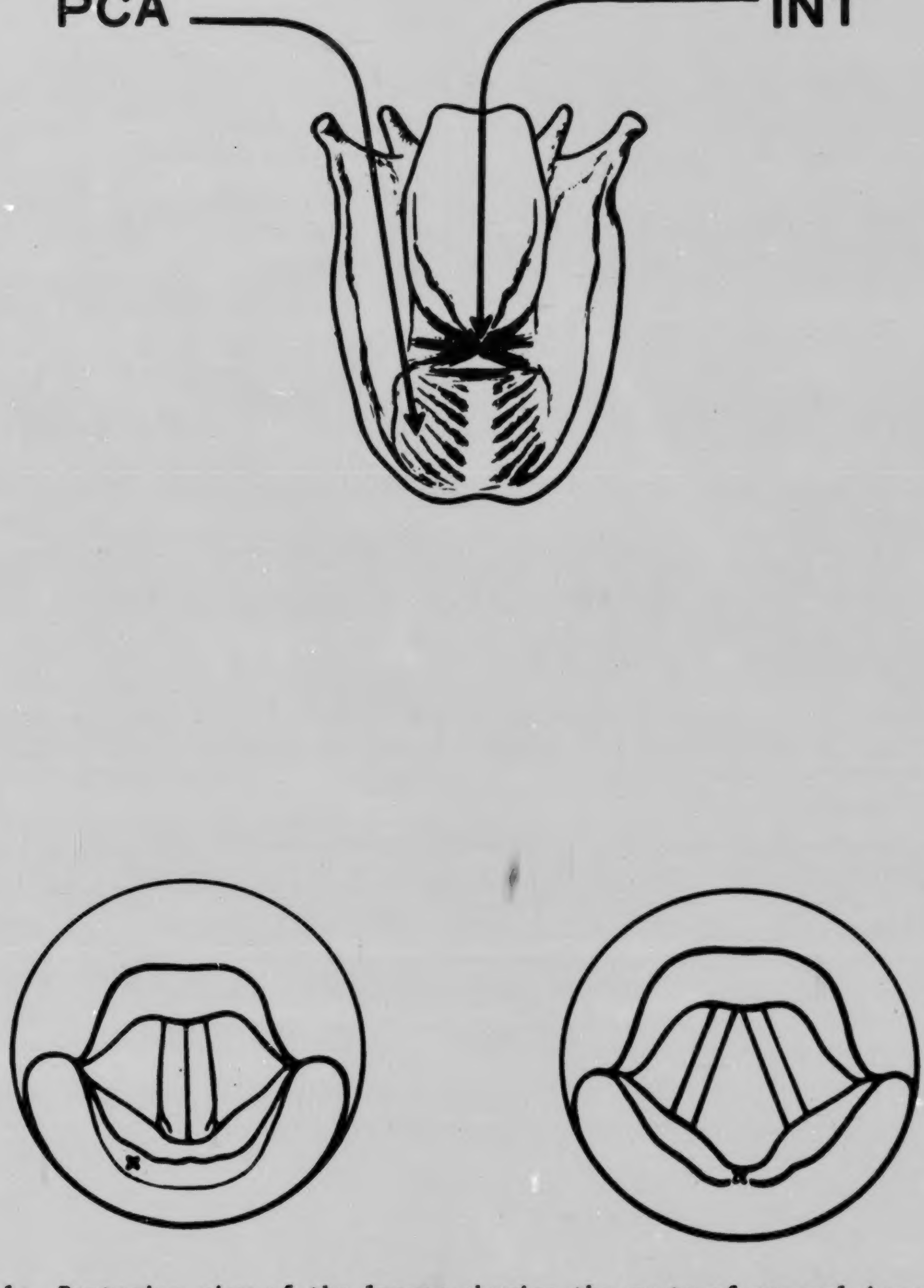


Figure 1: Posterior view of the larynx showing the route of peroral insertion of the electrodes into the PCA (left) and the INT (right).

Figure 2 shows an example of the averaged EMG data for the utterance "soreo seesee to yuu." The lower curve in the figure represents the averaged EMG activity of the PCA, while the upper curve represents that of the INT, which is shown for comparison. Zero on the time axis marks the onset of the acoustic signal for the vowel segment after the word-medial consonant.

Clearly, PCA activity is increasing for initial and medial /s/ in the test utterance and for /t/ in the carrier, while the INT shows reciprocal suppression for these voiceless segments. We also note that the peak value of PCA activity is almost the same for two /s/s in different positions.

Figure 3 illustrates the averaged EMG curves for the test utterance "soreo keekee to yuu." The lower curve again represents PCA activity. Although the PCA shows increasing activity for both /k/s in the test words, the peak value is higher for word-initial /k/ than for word-medial /k/.

In the second part of the experiment, high-speed motion pictures were taken of the glottis in the same subject and during the same test utterances for which EMG data were collected. The speaking rate was found to be consistent for both parts of the experiment. The motion pictures were taken through a fiberscope (Sawashima and Hirose, 1968) at a rate of 50 frames per second, and frame-by-frame analysis was performed.

For the voiceless portions of the test utterances, separation of the arytenoids and widening of the glottis were always observed. Figure 4 shows a comparison between the averaged time course of PCA activity (upper curve) and glottal width (lower curve) for the test word [ke:ke:], where glottal width was measured at the vocal process. It appears that the temporal course of glottal width is comparable to that of PCA activity--with some time delay. This holds true for all the utterance types examined.

Figure 5 compares a geminate and a devoiced vowel segment in word-medial position. Again, the time courses of PCA activity and glottal width are comparable. It should be noted in this figure that peak PCA activity is higher for the devoiced segment than for the geminate, although the duration of the glottal opening appears to be almost the same for these two. It has been reported that there is no systematic relationship between duration and maximum width of glottal opening (Sawashima and Miyazaki, 1973; Dixit and MacNeilage, 1974). The results of the present experiment are in good agreement with those previous reports.

Figure 6 illustrates the relationship between the peak values of averaged PCA activity and maximum glottal width for all types of voiceless segments used in the present experiment. As we can see, the maximum glottal width is generally larger when the peak activity is higher. A statistical test indicates that there is a significant positive correlation between these two parameters at the 0.001 level of confidence ($r = 0.86$).

Based on their fiberoptic observations, Sawashima and his colleagues have reported that the glottal opening for word-medial voiceless stops and geminates is generally smaller than that for voiceless fricatives or devoiced vowel segments in Japanese (Sawashima, 1971; Sawashima and Miyazaki, 1973). The present study supports their findings. Our results further indicate that the degree and timing of PCA activity are directly responsible for determining the size and temporal course of the glottal opening for voiceless segments, although the

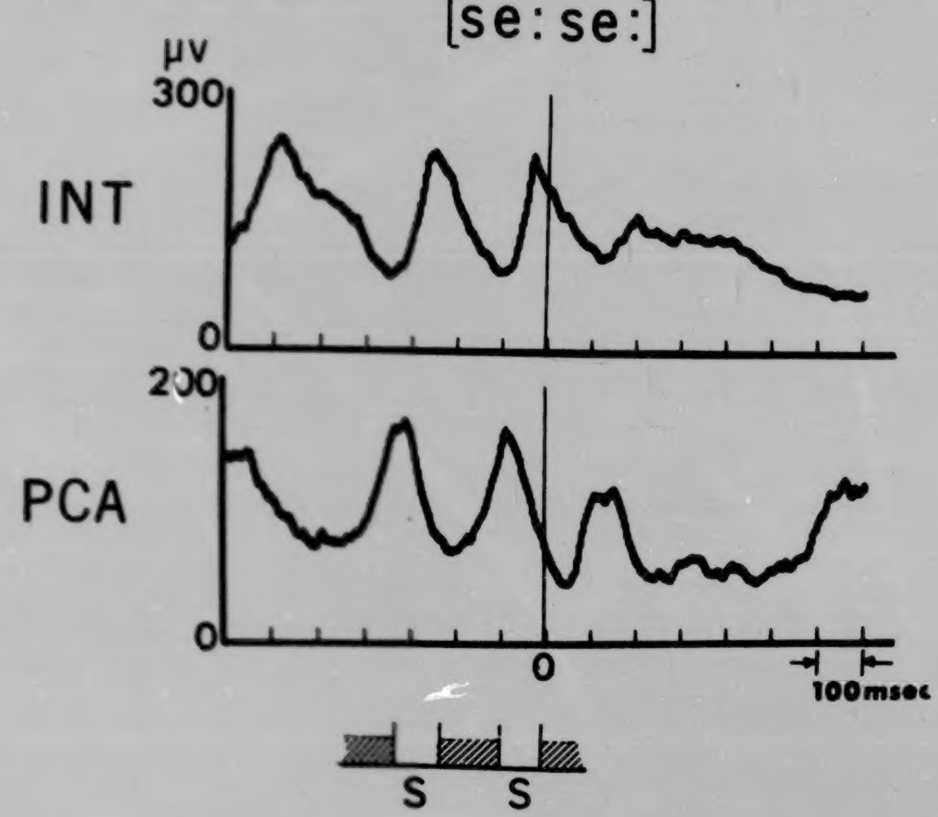


Figure 2: An example of the averaged EMG curves of INT (upper) and PCA (lower) for the test utterance "soreo seesee to yuu."

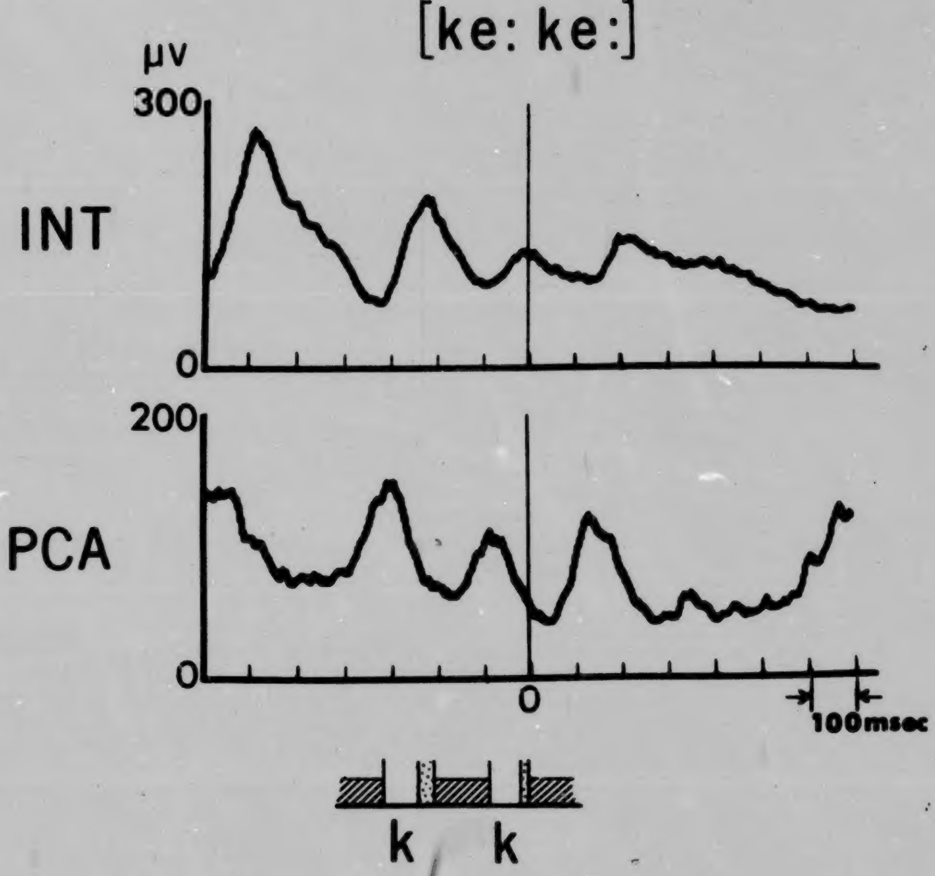


Figure 3: Averaged EMG curves of INT and PCA for the test utterance "soreo keekee to yuu."

FIGURE 4

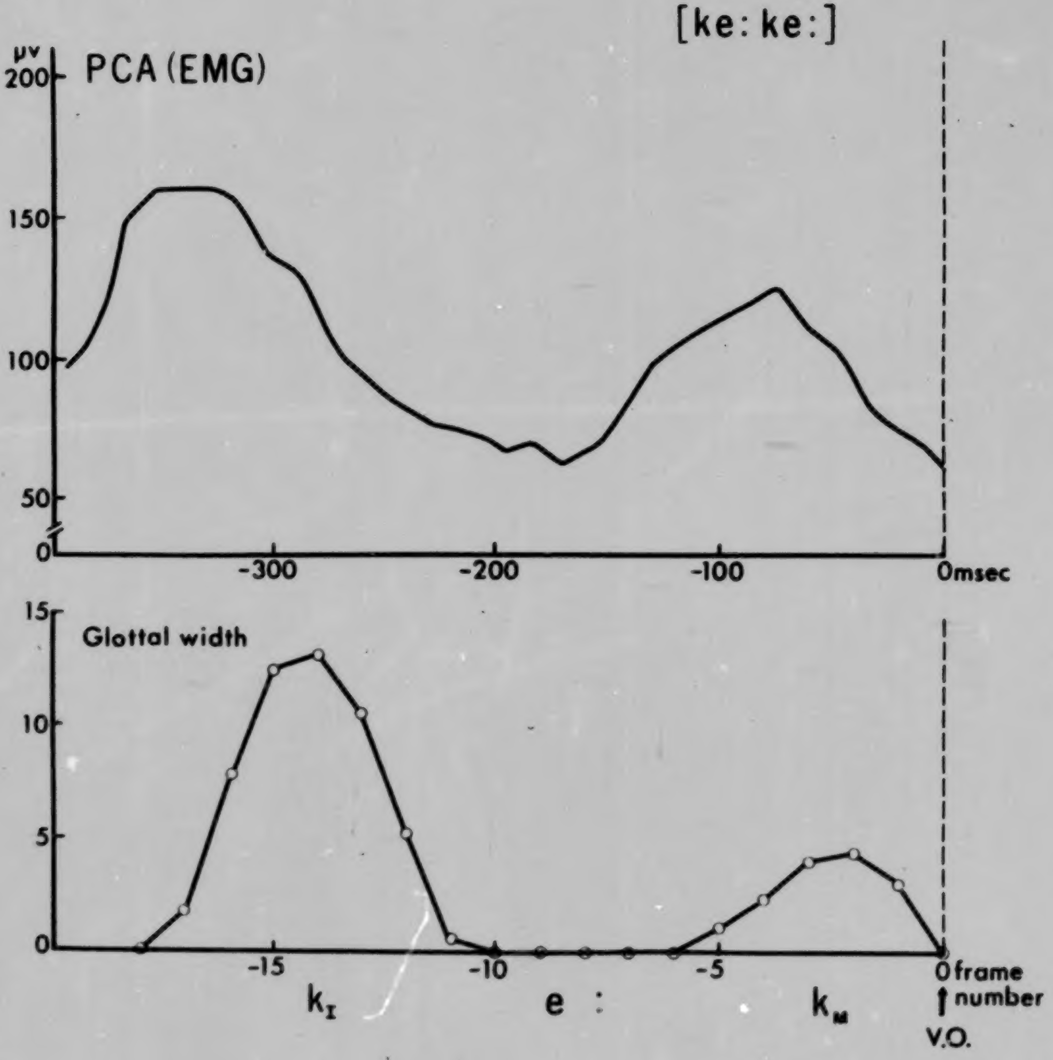


Figure 4: Comparison of the time course of PCA activity and glottal width for the test word [ke:ke:].

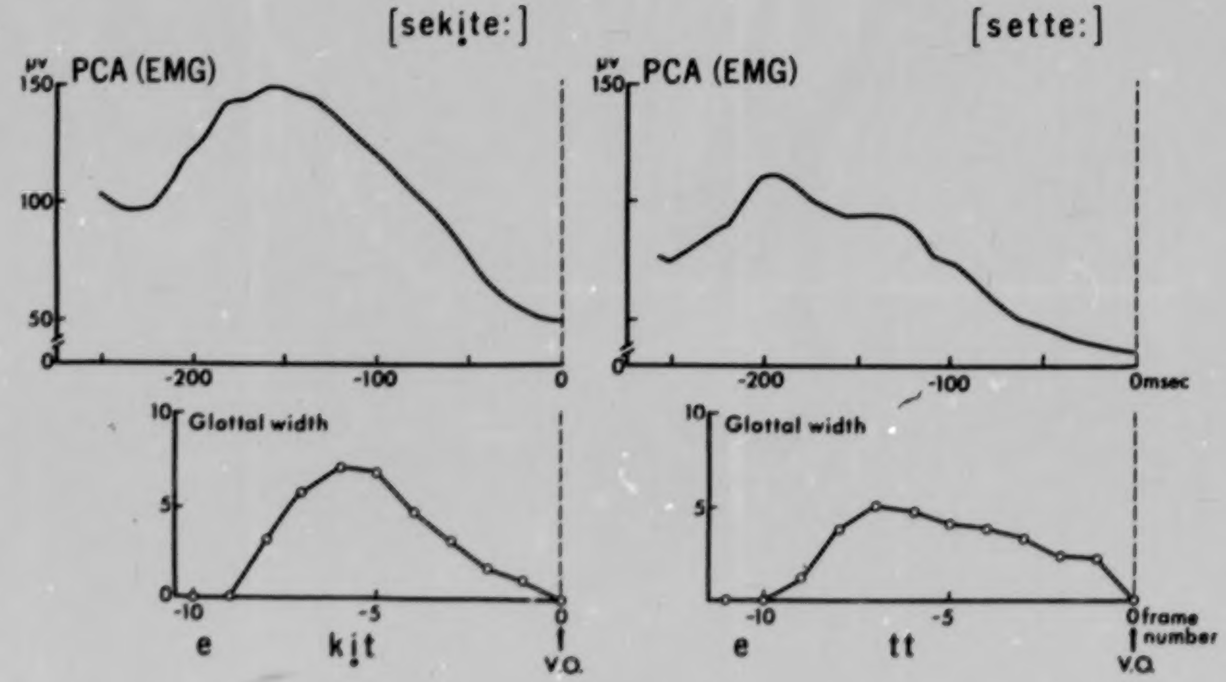


FIGURE 5

Figure 5: Comparison of the time course of PCA activity and glottal width for the word-medial voiceless segments [kịt] (left) and [tt] (right).

FIGURE 6

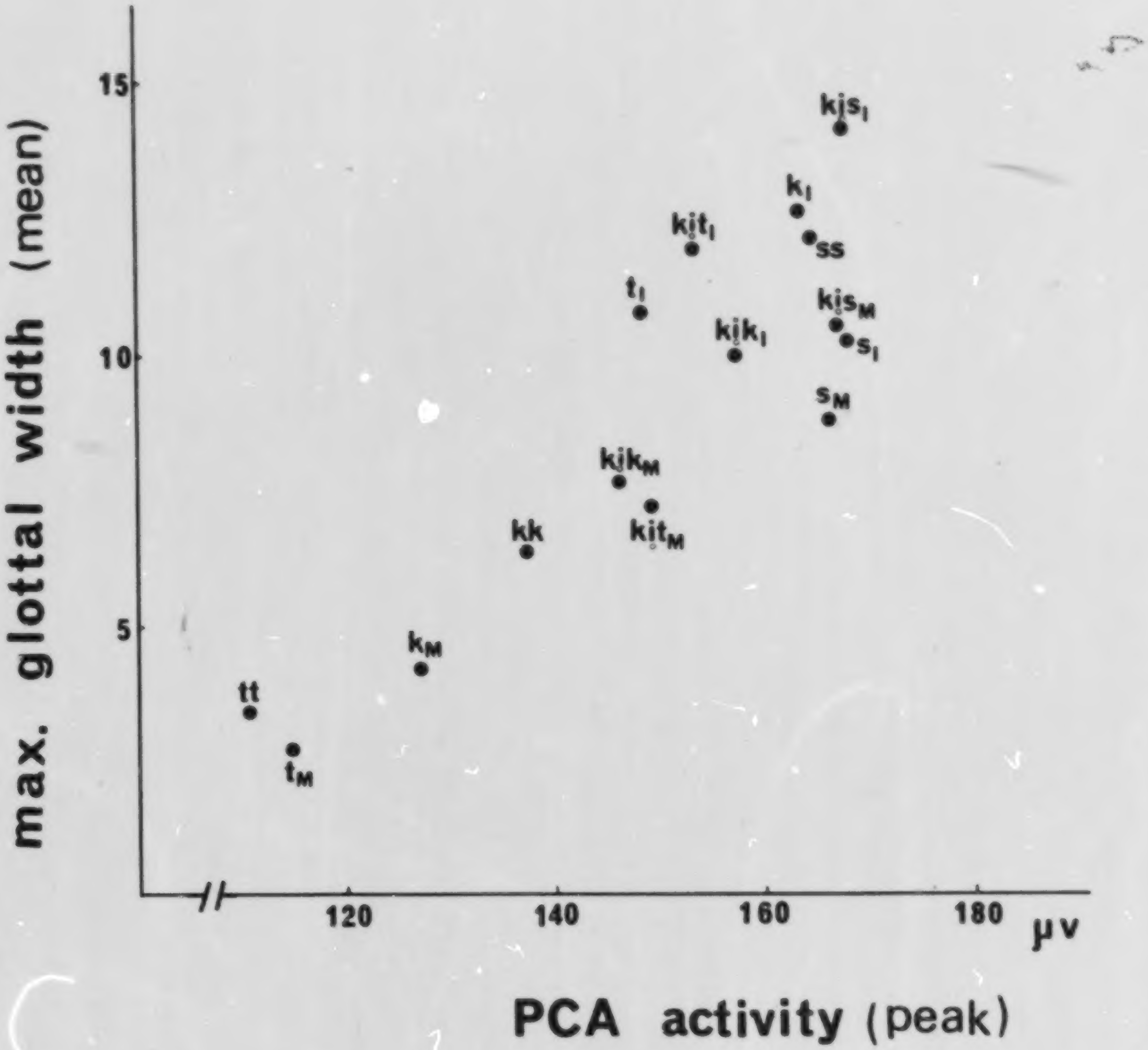


Figure 6: Relationship between the maximum glottal width (ordinate is an arbitrary scale) and peak value of averaged PCA activity (abscissa) for the voiceless segments examined in the present experiment.

suppression of the adductors may also have to be taken into consideration for a complete description of voiceless segment production.

REFERENCES

Dixit, R. P. and P. F. MacNeilage. (1974) Glottal dynamics during Hindi bilabial plosives and the glottal fricative. J. Acoust. Soc. Amer. 55 (Suppl.), S80.

Hirose, H. (1971) Laryngeal adjustments for vowel devoicing in Japanese. Haskins Laboratories Status Report on Speech Research SR-28, 157-166.

Hirose, H. (1973) The function of the posterior cricoarytenoid muscle--with special reference to laryngeal articulatory adjustments. Otologia (Fukuoka) 19, 711-723 [in Japanese].

Hirose, H. and T. Gay. (1972) The activity of the intrinsic laryngeal muscles in voicing control--an electromyographic study. Phonetica 25, 140-164.

Hirose, H., L. Lisker, and A. M. Abramson. (1972) Physiological aspects of certain laryngeal features in stop production. Haskins Laboratories Status Report on Speech Research SR-31/32, 183-191.

Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33 173-183.

Sawashima, M. (1971) Devoicing of vowels. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) 5, 7-13.

Sawashima, M. and H. Hirose. (1968) New laryngoscopic technique by use of fiberoptics. J. Acoust. Soc. Amer. 43, 168-169.

Sawashima, M. and S. Miyazaki. (1973) Glottal opening for Japanese voiceless consonants. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) 7, 1-9.

Laryngeal Activity Accompanying the Moment of Stuttering: A Preliminary Report of EMG Investigations*

Frances J. Freeman[+] and Tatsujiro Ushijima[++]
Haskins Laboratories, New Haven, Conn.

Throughout the history of man's interest in stuttering, certain conditions or circumstances have been found to produce immediate and marked improvement in the stutterer's fluency. As surveyed and reported by Bloodstein in 1950, such conditions include: (1) speaking to an imposed rhythm, (2) singing, (3) choral speaking and shadowing, (4) whispering, (5) shouting, (6) imitation of another speaker or dialect, and (7) speaking under conditions of diminished auditory sensitivity. Wingate (1969, 1970) has advanced the hypothesis that the conditions that effect a notable reduction in stuttering have a common feature--they all reflect some change in the mode or manner of vocalizing. According to Wingate, in these circumstances which improve fluency, "the stutterer does something with his voice that he does not ordinarily do."

Adams and Reis (1971) tested the relationship between phonation and dysfluency with an experiment using two 100-word prose passages. One passage was composed entirely of voiced sounds, while the other contained both voiced and voiceless segments. Stutterers had significantly fewer blocks in reading the all-voiced passage. More stuttering occurred when they had to make voiced-voiceless adjustments. The experiment has since been replicated with essentially the same results (Adams and Reis, in press).

Brenner, Perkins, and Soderberg (1972) looked at the effects of four rehearsal conditions on stuttering. They compared silent rehearsals without lip movement, silent rehearsals with lip movement, whispered rehearsals, and speaking aloud rehearsals. Only the rehearsal condition of speaking aloud resulted in significantly less stuttering. The authors concluded that stutterers have difficulty coordinating phonatory movements with articulatory movements.

These studies implicate the phonatory mechanism in stuttering by demonstrating changes in overt stuttering behavior--changes that result from manipulation of variables related to phonation.

---

*Paper presented at the 87th meeting of the Acoustical Society of America, New York, April 1974.

[+] Also City University of New York.

[++] Also University of Tokyo, Japan.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

Using techniques of direct and indirect observations, three studies have reported positive findings of laryngeal involvement in the moment of stuttering. Chevrie-Muller (1963) used a glottalgraphic technique with 27 stutterers; Fujita (1966) did a cinelaryngographic study of a Japanese stutterer; and Ushijima, Kamiyama, Hirose, and Niimi (1969) used the fiberscope to film laryngeal activity in stuttering.

The present electromyographic study attempts to move one step further into the speech production system to investigate the "motor commands" that result in the abnormal movement patterns observed by these researchers.

The experimental procedures are those developed at Haskins Laboratories and reported previously (Hirose, 1971; Port, 1971; Gay, Strome, Hirose, and Sawashima, 1972; Hirose and Gay, 1972, 1973; Kewley-Port, 1973, 1974). This preliminary paper reports data obtained on only one stuttering subject. Simultaneous recordings were obtained from four intrinsic laryngeal muscles (the posterior cricoarytenoid, the lateral cricoarytenoid, the vocalis, and the cricothyroid); three lingual muscles (the inferior longitudinal, the superior longitudinal, and the genioglossus); and one labial muscle (the orbicularis oris).

Comparisons were made of the stuttering subject's fluent and stuttered utterances of the same words. Similar comparisons were made of a normal speaking subject's fluent and "faked" stuttered utterances. Results indicate that fluent utterance is characterized by precise balance and timing of laryngeal abductor and adductor forces.

Figure 1 illustrates the normal pattern of adductor-abductor forces. In this pattern, increases in abductor activity accompany decreases in adductor activity, and conversely, when adductive activity increases abductor activity decreases. Activity patterns are shown here for three intrinsic laryngeal muscles--the posterior cricoarytenoid, the vocalis, and the lateral cricoarytenoid. The top tracing is for the abductor and the two lower are for adductors. On the left is an averaged number of tokens of the utterance glottal stop /a/, while on the right is a single token record for a swallow. All of the stuttered data represent single tokens.

Note that the lateral cricoarytenoid, an adductor with the specific function of applying medial compression, shows a high level of activity for the tight closure of the glottal stop and for the first portion of the swallow. As an adductor, the vocalis participates in the glottal stop and swallow closures, and like the lateral cricoarytenoid, is active for the vowel segment. The posterior cricoarytenoid (PCA) is suppressed during the closure periods of the glottal stop and the swallow and also during the vowel segment. A brief, very slight increase in PCA activity occurs just after the glottal stop, and a strong burst of PCA activity follows the closure in the swallow. The abductor-adductor reciprocity is readily apparent.

The abductor-adductor reciprocity, so characteristic of fluent utterance, is disrupted in stuttered utterance. The most common pattern occurring during moments of stuttering is simultaneous, presumably antagonistic, abductor-adductor activity.

In Figure 2 the stuttered utterance of the word "llllllless" is contrasted with the fluent utterance "less." The upper channel traces the abductor activity

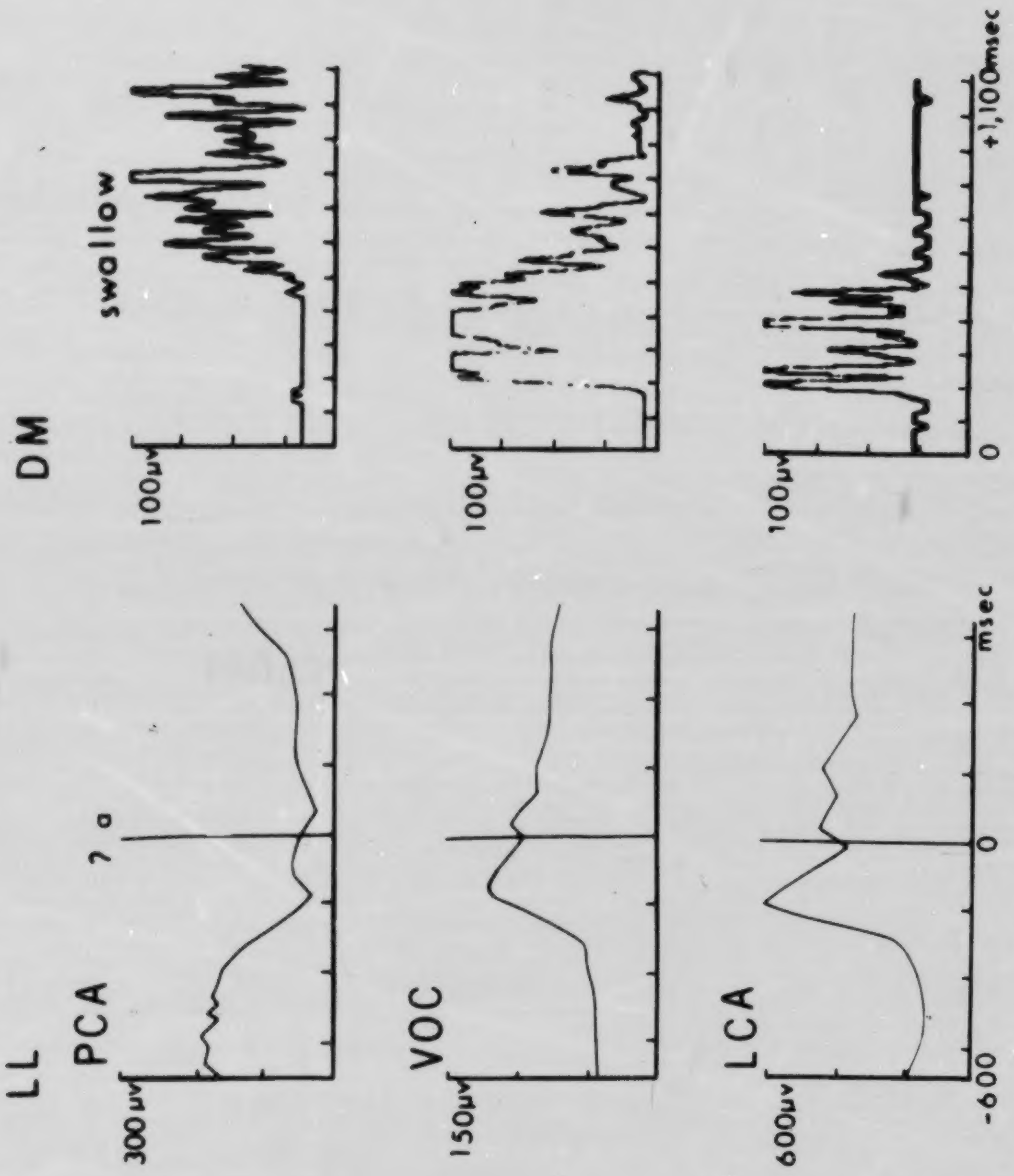

LL
PCA
VOC
LCA
300µv
150µv
600µv
? a
-600
0
msec
DM
swallow
100µv
100µv
100µv
0
+1,100msec

FIGURE 1

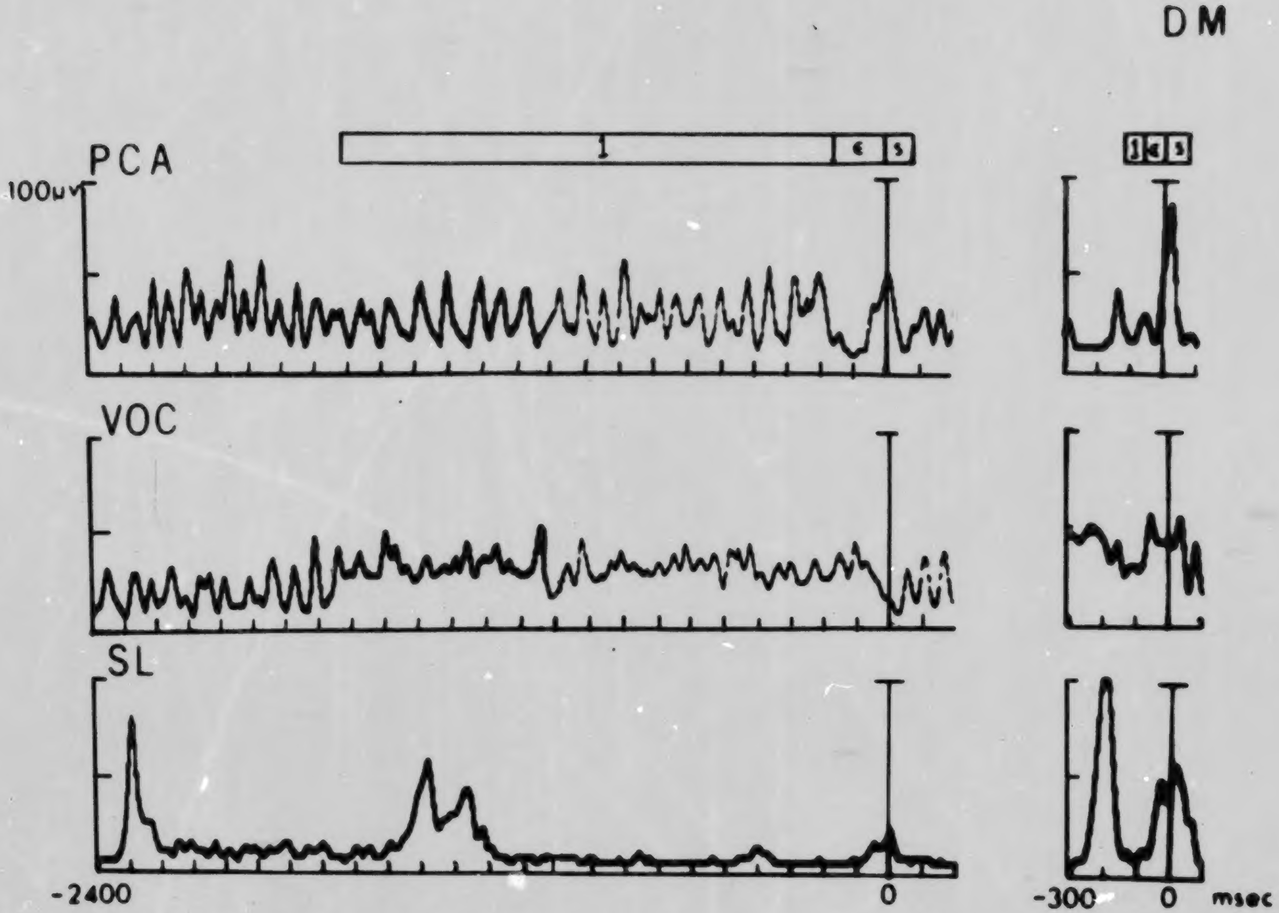


FIGURE 2

of the PCA, while the second channel shows the adductor activity of the vocalis. The third channel is the superior longitudinal. Activity in this muscle correlates with raising and retraction of the tongue tip. Here activity in the superior longitudinal, presumably for raising the tongue tip for /l/, occurs 2300 msec before the word is uttered. This activity occurs 660 msec before any acoustic signal is detected. During this "silent" period, PCA abductive activity gradually builds, as does activity in the vocalis. At this point (-1600 msec) a higher level of vocalis activity corresponds to the onset of the prolonged utterance of the /l/. The segment is sustained for 1420 msec, a period characterized by high levels of simultaneous adductor-abductor activity. One-hundred-sixty msec before the lineup point, two things occur: there is a sharp drop in PCA activity, and the subject moves through the block. Reciprocity appears reestablished, for the following increase in PCA activity for the voiceless segment /s/ is timed to correspond to a marked suppression of adductor activity. The fluent utterance shown on the right of Figure 2 requires less than 300 msec.

In addition to the disruptions of reciprocity already noted, the stuttered utterances were frequently characterized by high levels of lateral cricoarytenoid activity. In direct contrast, the subject's periods of fluent utterance were found to occur in association with marked suppression of activity in this adductor. Figure 3 illustrates this finding. In this example, the stutterer uttered the word "effect" three times, with progressive adaptation from a severe block to a mild block to a fluent utterance. The degree of lateral cricoarytenoid activity correlates with the degree of dysfluency.

In most cases, higher levels of activity were recorded during stuttering blocks than during fluent utterance. The successful termination of a block was frequently found to coincide with a marked drop in adductor and/or abductor activity.

Figure 4, which shows progressive adaptation in the utterance of the word "ancient," illustrates each of the three findings already discussed:

1) Abductor-adductor reciprocity is disrupted in the two stuttered utterances;

2) Progressively lower levels of lateral cricoarytenoid activity accompany the more fluent utterances; and

3) Somewhat higher levels of activity occur in the strongly stuttered utterance, where a marked drop in activity coincides with the termination of the block.

In fact, inspection of the final portions of the stuttered utterances indicates that successful termination of a block coincides with a pattern of laryngeal muscle activity that approximates the pattern characteristic of fluent utterance of the same word. In other words, the same balance of abductor-adductor forces characteristic of fluent utterance is characteristic of the termination of the block.

REFERENCES

Adams, M. and R. Reis. (1971) The influence of the onset of phonation on the frequency of stuttering. J. Speech Hearing Res. 14, 639-644.

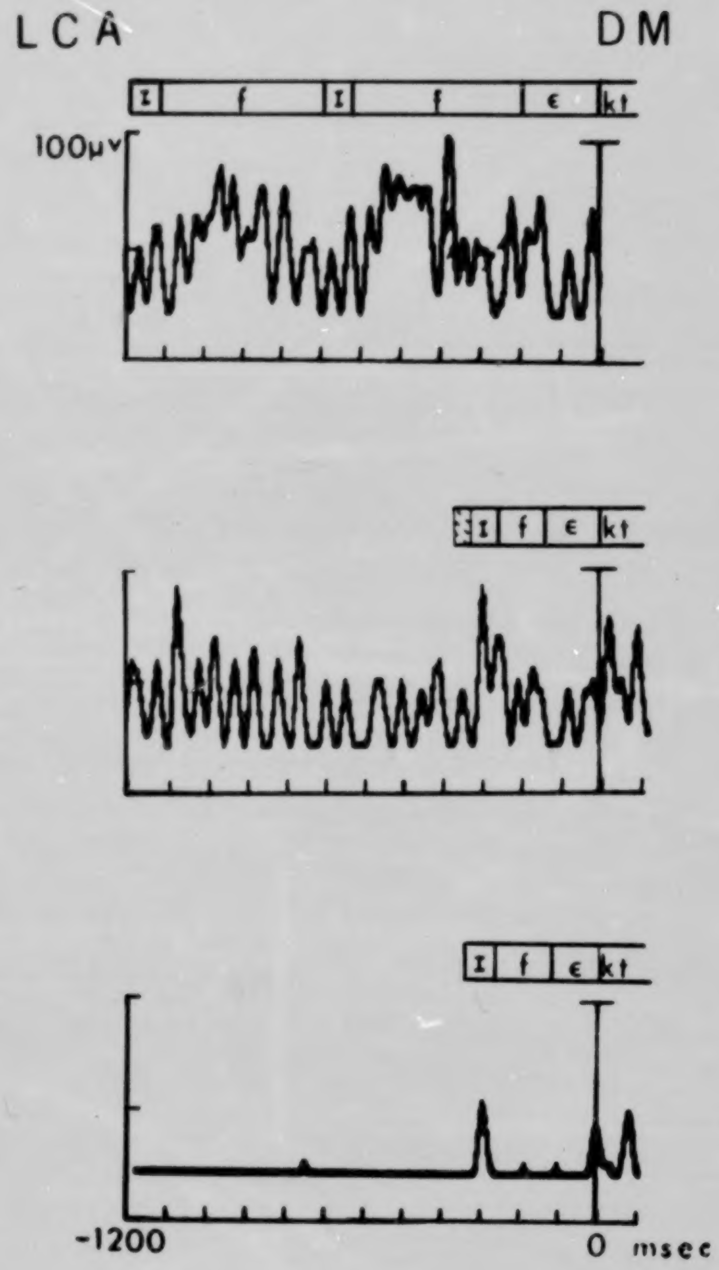


FIGURE 3

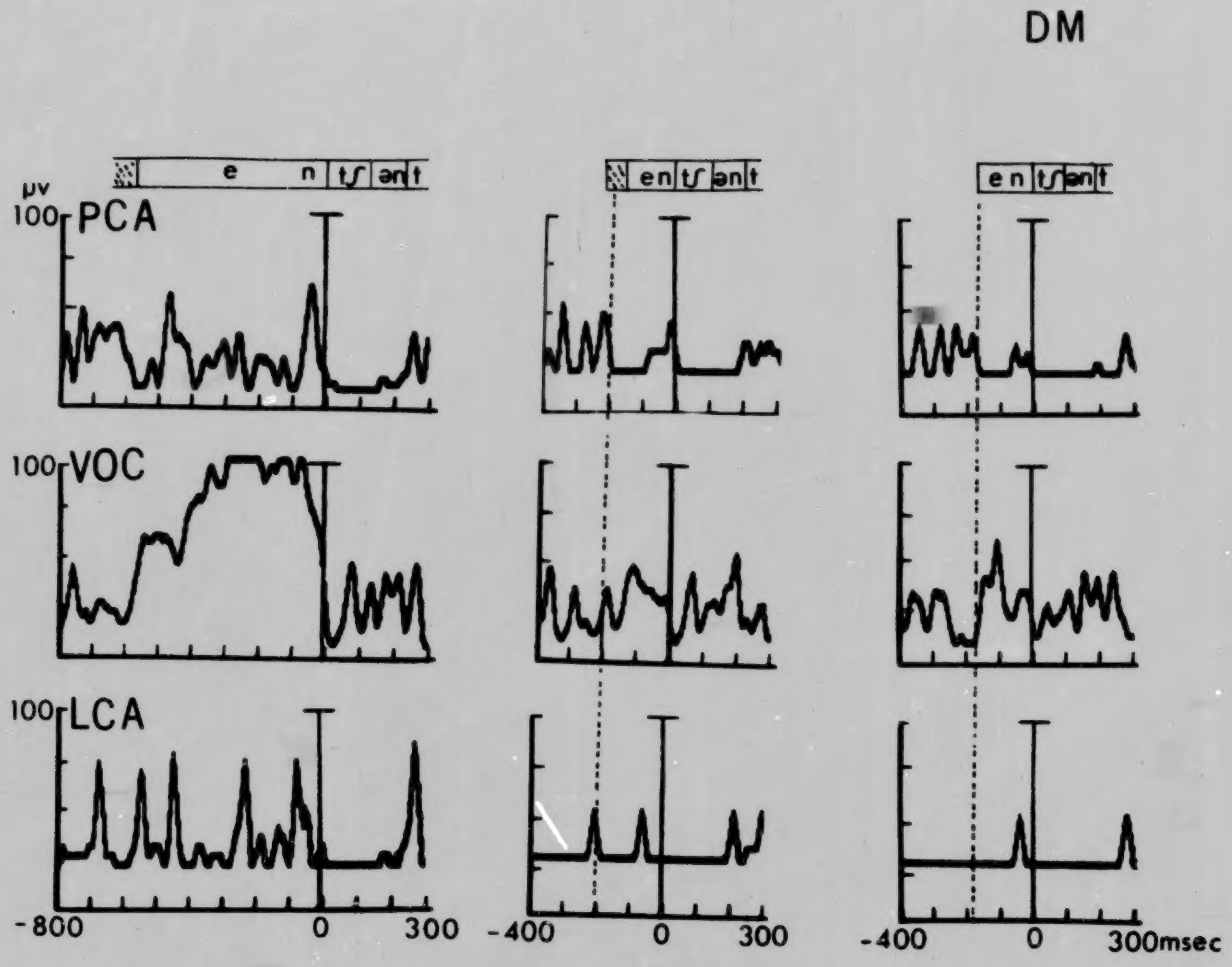

DM
e n tʃ ənt
en tʃ ənt
en tʃ ənt
μv
100 PCA
100 VOC
100 LCA
-800
0
300
-400
0
300
-400
0
300msec

FIGURE 4

Adams, M. and R. Reis. (in press) The influence of the onset of phonation on the frequency of stuttering: A replication and reevaluation. J. Speech Hearing Res.

Bloodstein, O. (1950) A rating scale study of conditions under which stuttering is reduced or absent. J. Speech Hearing Dis. 15, 19-36.

Brenner, N. C., W. H. Perkins, and G. A. Soderberg. (1972) The effect of rehearsal on frequency of stuttering. J. Speech Hearing Res. 15, 474-482.

Chevrie-Muller, C. (1963) A study of laryngeal function in stutterers by the glottalgraphic method. In Proc. VII Congress de la Societe Francaise de Medecine de la Voix at de la Parole, Paris.

Fujita, K. (1966) Pathophysiology of the larynx from the viewpoint of phonation. J. Japan. Soc. Otorhinolaryngol. 69, 459.

Gay, T., M. Strome, H. Hirose, and M. Sawashima. (1972) Electromyography of the intrinsic laryngeal muscles during phonation. Ann. Otol., Rhinol., Laryngol. 81, 401-408.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Hirose, H. and T. Gay. (1972) The activity of the intrinsic laryngeal muscles in voicing control: An electromyographic study. Phonetica 25, 140-164.

Hirose, H. and T. Gay. (1973) Laryngeal control in vocal attack: An electromyographic study. Folia Phoniat. 25, 203-213.

Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-184.

Kewley-Port, D. (1974) An experimental evaluation of the EMG data processing system: Time constant choice for digital integration. Haskins Laboratories Status Report on Speech Research SR-37/38 (this issue).

Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

Ushijima, T., G. Kamiyama, H. Hirose, and S. Niimi. (1969) Articulatory Movements of the Larynx in Stuttering. Filmed at the University of Tokyo, Japan.

Wingate, M. E. (1969) Sound and pattern in "artificial" fluency. J. Speech Hearing Res. 12, 677-686.

Wingate, M. E. (1970) Effect on stuttering of changes in audition. J. Speech Hearing Res. 13, 861-873.

Hemispheric Lateralization for Speech Perception in Stutterers

M. F. Dorman[+] and R. J. Porter, Jr.[++]

Some authors have suggested that stutterers suffer from incomplete cerebral lateralization for speech (Orton, 1928; Travis, 1931; Beech and Fransella, 1968). In this view, often called the Orton-Travis theory, an absence of normal cerebral dominance is thought to result in an incoordination of cortical areas underlying speech production and perception. Early attempts to test this possibility (Bryngelson, 1935, 1940; Heltman, 1940) were inconclusive, perhaps due to the inherently low reliability of the measures of cerebral lateralization employed (e.g., handedness). Renewed interest in testing the theory has developed, however, because of a new, and possibly more reliable, behavioral measure of cerebral lateralization of auditory function introduced by Kimura (1961a, 1961b).

Several varieties of Kimura's task now exist (Berlin and McNeil, in press). However, all share a common component. Subjects are asked to identify and/or recall contrasting pairs of speech sounds, each member of the pair being presented to a different ear. Under such dichotic competition, subjects tend to report the right-ear stimuli more accurately than the left-ear stimuli. This right-ear advantage (REA) can be interpreted as reflecting the left-hemisphere's specialization for speech and language processing (Kimura, 1961b; Studdert-Kennedy and Shankweiler, 1970; Berlin, Lowe-Bell, Cullen, Thompson, and Loovis, 1973).

Several investigators have attempted to test the Orton-Travis theory by administering dichotic listening tasks to stutterers and nonstuttering control subjects. The results have been contradictory. Curry and Gregory (1969) found support for the Orton-Travis theory when a majority of the stutterers they tested evidenced better left- than right-ear report on a dichotic word task. In another test, Jones (1966), using the Wada intracarotid sodium amytal test (Wada and Rasmussen, 1960), found bilateral speech representation in four stutterers who underwent surgery for brain injury. Quinn (1972), however, has reported no

---

[+] Haskins Laboratories, New Haven, Conn., and Lehman College of the City University of New York.

[++] Department of Psychology, University of New Orleans, La., and Kresge Hearing Research Laboratory of the South, Louisiana State University Medical Center.

Acknowledgment: This study was conducted while the authors were at the University of Connecticut, Storrs. The assistance of Dr. R. L. Webster, the staff of the Institute for Behavioral Research, and the Departments of Psychology and Linguistics at the University of Connecticut is gratefully acknowledged. The research was supported, in part, by NIH Pre-Doctoral fellowship 1-F01-MH-45, 975-01MTLH to the second author, and by a grant to Haskins Laboratories.

differences between adult stutterers and controls on a dichotic listening task, and Slorach and Noehr (1973) have obtained similarly negative results with six- to nine-year-old stutterers and controls. This discrepancy in dichotic results may be due, in part, to the fairly large variability in REAs obtained with some dichotic tasks (Porter, in press) and to the difficulty of obtaining samples of stutterers' homogeneous in handedness, degree of speech impairment, etc.

In the present study, adult, right-handed, moderate-to-severe stutterers and normal-speaking subjects were presented a highly reliable dichotic nonsense-syllable task in order to probe further the possible relationship between hemispheric lateralization for speech and stuttering.

METHOD

Subjects

The subjects were 16 right-handed, adult stutterers (12 males, 4 females) and 20 nonstutterers (10 males, 10 females). The stutterers were drawn from therapy programs at the Institute for Behavioral Research (Summer, 1969) and the University of Connecticut (1970). All were moderate-to-severe stutterers with at least a 10-year history of stuttering. The nonstutterers, students at the University of Connecticut, were given class credit for participation. All subjects had normal hearing (by self report) and were native speakers of American English.

Preparation of Stimuli

Synthetic signals appropriate for consonant-vowel syllables [ba, da, ga, pa, ta, ka] were generated with the aid of the Haskins Laboratories' speech synthesizer. Under computer control these six stimuli were combined into the 15 possible contrasting pairs and were recorded dichotically in a fully counterbalanced, random order onto magnetic tape. The resulting tape contained 60 stimulus pairs with each member of a pair occurring twice on each channel. The interpair interval was 4 sec. The stimuli were reproduced on an Ampex AG 500 or a General Radio tape recorder and presented via matched TDH-39 headphones. The outputs of the tape channels were equated (within 1 db) and monitored by voltmeter. The signal level was 75 db SPL ±5 db.

Procedure

In order to familiarize the subjects with the stimuli, and to discover any gross hearing deficits, the subjects were first presented two monaural syllable identification tests (one to each ear). (All subjects performed at virtually 100% on these monaural tasks.) Before dichotic testing the subjects were told they would hear two syllables simultaneously and were instructed to write the identity of both syllables, in order of clarity, on an answer sheet. The subjects were given three dichotic practice trials followed by two presentations of the 60-item dichotic test. The subjects' headphones were reversed for the second 60-item test in order to counterbalance any channel imbalances.

RESULTS

The mean number of dichotic syllables correctly reported (maximum of 120 for each ear) from the right and left ears for both stutterers and controls, subcategorized by sex, is shown in Table 1. Significant REAs were found for both male stutterers and male controls. The magnitude of the REAs did not differ between

TABLE 1: Mean number of syllables correctly reported from each ear.

Group		N	Left	Right	t
Stutterers:	M	12	20.25	27.04	2.21*
	F	4	17.51	29.62	2.16
Controls:	M	10	19.80	26.15	2.92*
	F	10	15.52	34.71	5.91**

*$p < .05$
**$p < .01$

these groups ($t_{20} = 0.147$, $p > .05$). A significant REA was also found for the female controls. Three of the four female stutterers evidenced large REAs (S_1 = 38%; S_2 = 52%; S_3 = 13%; S_4 = 0%), but the REA was not significant. Because of the small number of female stutterers, the statistical analysis of these data and a comparison with controls must be made with some caution. The female stutterers' results do, however, fall within the range of the control results, and there appears to be no reason to classify them as abnormal.

Within the control population, females evidenced a significantly larger REA than males ($t_9 = 3.55$, $p < .01$). The mean scores for male and female stutterers bear the same relation as those for male and female controls.

A summary of the findings in terms of the metric $\frac{R-L}{R+L} \times 100$, where R (or L) is the number of syllables correctly reported from the right (or left) ear, is shown in Table 2.

TABLE 2: Mean ear advantage (%) in terms of $\frac{R-L}{R+L} \times 100$.

	Stutterers	Controls
Males	14.78 n=12	13.81 n=10
Females	25.65 n=4	38.20 n=10

DISCUSSION

Both male and female stutterers identified syllables presented to the right ear better than syllables presented to the left ear. Furthermore, the magnitude of the REA for the stutterers as a group was very similar to that of the controls as a group. Clearly, these data fail to lend support to the theory that stutterers suffer abnormalities in speech lateralization.

Although the absolute magnitude of the female stutterers' REA was smaller than that of the female controls' REA, all stutterers' REAs were well within the range of REAs found in normal populations (Studdert-Kennedy and Shankweiler, 1972). In fact, if any group performance approaches the extremes of the normal population, it is that of the female control group.

In summary, the present data, those of Quinn (1972), and those of Slorach and Noehr (1973) indicate that stutterers fall well within the normal range of lateralization for speech as indicated by a dichotic test. Since it has also been demonstrated that individuals with bilateral speech representation (as determined by the Wada test) may have normal speech ability (Milner, Branch, and Rasmussen, 1964), it would appear that factors other than abnormalities in cortical lateralization underlie stuttering.

SUMMARY

Sixteen adult, right-handed, moderate-to-severe stutterers (12 males, 4 females) and 20 nonstuttering controls (10 males, 10 females) were given a dichotic nonsense-syllable test to determine hemispheric lateralization for speech. Both male and female stutterers evidenced right-ear advantages in syllable identification similar in magnitude to those found for normals. These data confirm other reports of no difference in cerebral speech lateralization for stutterers and nonstutterers and, therefore, lend no support to theories that relate stuttering to abnormalities in cerebral lateralization.

REFERENCES

Beech, H. R. and F. Fransella. (1968) Research and Experiment in Stuttering. (Oxford: Pergamon).

Berlin, C. I., S. S. Lowe-Bell, J. K. Cullen, C. L. Thompson, and C. F. Loovis. (1973) Dichotic speech perception: An interpretation of right-ear advantage and temporal offset effects. J. Acoust. Soc. Amer. 53, 699-709.

Berlin, C. I. and M. R. McNeil. (in press) Dichotic listening. In Contemporary Issues in Experimental Phonetics, ed. by Norman J. Lass. (Springfield, Ill.: Charles C Thomas).

Bryngelson, B. (1935) Sidedness as an etiological factor in stuttering. Pedagog. Semin. 47, 204-217.

Bryngelson, B. (1940) A study of laterality of stutterers and normal speakers. J. Soc. Psychol. 11, 151-155.

Curry, F. W. and H. H. Gregory. (1969) The performance of stutterers on a dichotic listening task thought to reflect cerebral dominance. J. Speech Hearing Res. 12, 73-82.

Heltman, H. J. (1940) Contradictory evidence in handedness and stuttering. J. Speech Dis. 5, 327-331.

Jones, R. K. (1966) Observations on stuttering after localized cerebral injury. J. Neurol. Neurosurg. Psychiat. 29, 192-195.

Kimura, D. (1961a) Some effects of temporal lobe damage on auditory perception. Canad. J. Psychol. 15, 156-165.

Kimura, D. (1961b) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Milner, B., C. Branch, and T. Rasmussen. (1964) Observations on cerebral dominance. In Disorders of Language: A CIBA Foundation Symposium, ed. by A. V. S. DeReuck and M. O. O'Connor. (London: Churchill).

Orton, S. T. (1928) Studies in stuttering. Arch. Neurol. Psychiat. 18, 671-672.

Porter, R. J. (in press) On interpreting developmental changes in the dichotic right-ear advantage. Brain and Language.

Quinn, P. (1972) Stuttering, cerebral dominance, and the dichotic word test. Med. J. Aust. 2, 639-643.

Slorach, N. and B. Noehr. (1973) Dichotic listening in stuttering and dyslalic children. Cortex 9, 295-300.

Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Studdert-Kennedy, M. and D. Shankweiler. (1972) A continuum of cerebral dominance for speech perception? Haskins Laboratories Status Report on Speech Research SR-31/32, 23-40.

Travis, L. E. (1931) Speech Pathology. (New York: Appleton-Century-Crofts).

Wada, J. and T. Rasmussen. (1960) Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance: Experimental and clinical observations. J. Neurosurg. 17, 266-282.

Dichotic Release from Masking: Further Results from Studies with Synthetic Speech Stimuli

P. W. Nye, T. M. Nearey,[+] and T. C. Rand
Haskins Laboratories, New Haven, Conn.

INTRODUCTION

The strong tendency of monaural low-frequency tones to mask high-frequency tones entering the same ear is a well-known phenomenon termed the "upward spread of masking." Average energy in the speech spectrum peaks in the low-frequency region (below 1000 Hz) occupied by the first formant and declines toward the higher frequencies at a rate of approximately 6 db per octave (Fant, 1950). Hence, the conditions exist for the lower band of energy, which conveys (for the most part) manner of articulation distinctions, to mask information in the upper frequencies, which broadly contain most of the information relating to place distinctions. A preliminary study at Haskins Laboratories, recently reported by Rand (1974), has indicated that the first formant (F_1) of a stop-vowel syllable can, under certain circumstances, mask the higher formants F_2 and F_3. The effect was first demonstrated by presenting speech signals dichotically (F_1 to one ear and F_2 and F_3 to the other ear) whereupon, Rand showed, a 20 db release from masking can occur.

The central purpose of this study is to determine whether there are conditions in which the dichotic release from masking can be exploited in speech communication. From this central issue three related questions emerge and these have provided the focus of work reported here. The questions in order are:

1) The existence question: Is there a release from masking? The replication of the original result proved to be more difficult than was first anticipated and a search was made to find the optimum conditions necessary to demonstrate the effects.

2) The noise question: Is the release from masking effect observed in conditions of added gaussian noise?

3) The phonetic range question: Does F_1 masking affect a broad range of phonetic material? That is, are English words and sentence-like strings affected in the same way as are nonsense stop + vowel syllables?

In answer to the first of these questions, the data broadly corroborate the existence of a strong release from masking in conditions of upper formant

---

[+] Also Department of Linguistics, University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

attenuation, provided the voiced portions of the signals exceed 70 db SPL under good listening conditions (S/N > 40 db). Furthermore, question 3 can be answered in the affirmative since word recognition scores remain higher under dichotic conditions when F_2 and F_3 are attenuated; also, the dichotic release from masking improves the recognition scores on a wide repertoire of phones. Question 2 can be answered, however, only with some qualification. The observations made so far indicate that a much smaller release from masking is obtained for signals in noise, and only when F_2 and F_3 are attenuated. No release has been demonstrated in noise when F_2 and F_3 are set at their natural energy levels relative to F_1.

In addition to these three topics, there is a fourth background issue of major significance. This involves the question of fusion. Because of the redundancy of phonetic information in the speech signal (a possibly acute problem in synthetic speech where certain redundancies may be exaggerated), it is no easy matter to determine whether, or to what extent, the listener combines acoustic information from both ears in the dichotic condition. This issue will be discussed in more detail in a concluding section.

The present paper represents an interim report of an ongoing research study. More work remains: specifically, a more detailed analysis of articulation scores on synthetic speech, a further examination of the effects of noise on the perception of dichotic speech (with F_2, F_3 unattenuated), an extension of the study to natural speech (digitally filtered and presented dichotically), and an investigation of the fusion question.

METHODS AND PROCEDURE

The stimuli used in the experiments discussed in this report were generated by a parallel formant resonance synthesizer designed and built at Haskins Laboratories. Using this instrument the formant resonances could be easily separated and recorded individually. The relative intensities of the formants were set at values consistent with the ratio measurements made by Fant (1950) on the vowel /ɑ/ in natural speech at a sound pressure level of 70 db. Formants F_2 and F_3 were recorded on channel A of a two-channel tape recorder--the other channel B being used to record F_1. For a variety of practical purposes it is of importance to learn whether the release from masking available under dichotic conditions will assist the listener in assimilating speech in noise. In conducting experiments to determine whether listening performance in noise can be improved, the spectrum of the added noise was made essentially flat over the speech bandwidth (60-6000 Hz) and zero elsewhere. Moveover, from a practical point of view the act of separating the first formant from the remainder of the speech signal tacitly assumes that a dynamically adjustable filter must be available to divide the speech spectrum automatically midway between the first and second formants. Such a filter would also divide the noise spectrum into two parts (in the region of 1 kHz) and in these experiments the added noise was divided in half in a corresponding fashion. In most cases where noise was required it was introduced during the recording process and the relative signal/noise levels were checked on replay. Alternatively, noise was supplied at replay time by a noise generator or by a prefiltered, prerecorded tape. The stimulus recording level was standardized at -1 db VU measured on the sustained vowel /ɑ/ in the case of the syllables and on the vowel /æ/ for word and sentence stimuli. On replay the stimuli were heard by the experimental subjects via Grason-Stadler earphones type TDH39-300Z. Measurements performed on these earphones, using a sound pressure meter and artificial ear coupler, provided a calibration curve relating RMS signal voltages (input to an earphone) to sound pressure level in decibels delivered to the ear drum. The standard signal for these measurements was again the sustained vowel /ɑ/. Each recording of stimulus material included a passage of sustained /ɑ/ output with which it was possible to adjust the gain of reproducing amplifiers to achieve any desired RMS input to the earphones and hence determine the sound pressure impinging on the subjects' ears. We refer to a level adjusted in this manner as the "baseline SPL": it constitutes the total energy of all three formants received by one or both ears when listening under monaural or binaural conditions. In dichotic conditions, the formants are separated and the SPL at any single ear drum is dependent upon the particular portion of the signal being transmitted. In general, therefore, this level is always some fraction of the baseline level. In the case of dichotic /ɑ/ the true sound pressure level of the first formant was 1.5 db below the "baseline SPL," and formants F_2 and F_3 lay 3 db below the nominal or baseline level.

The majority of the listeners in the experiments were college students 20 to 25 years of age with good hearing. In all cases where a separate ear analysis was to be made, auditory sensitivity was checked with a screening instrument, as it was on other occasions when there was reason to suspect that a subject's sensitivity was below normal limits. However, it was found that all subjects with a known hearing imbalance (usually < 10 db) produced data that were statistically indistinguishable from the data of their peers and these data were therefore included in the final analysis.

RESULTS

Experiment 1

This experiment was conducted in an effort to find the optimum conditions for a strong release from masking. Designed in three stages, the experiment examined the differences among binaural, monaural, and dichotic listening performances as a function of (a) the baseline sound pressure level, (b) the signal/noise ratio, and (c) the attenuation of F_2 and F_3. Nine subjects took part in all three stages and were required to identify the syllables [ba], [da], and [ga]. In stage (a) the baseline stimulus intensity was varied in 10-db steps between 60 and 90 db SPL. At each baseline level the F_2, F_3 signal was attenuated by a constant 30 db. Stage (b) utilized stimuli having a baseline sound pressure level of 80 db with the F_2, F_3 signals attenuated by a constant 20 db. To these signals was added random gaussian noise band-split at 1 kHz--the lower band being mixed with F_1 and the upper band with F_2, F_3. Monaural stimuli were recorded with full band noise. Four signal/noise ratios were examined: +12 db, +17 db, +24 db, and +45 db (the latter being the nominal limit for the recording and reproducing system in the absence of any externally applied noise source). Stage (c) employed the syllables at a baseline level of 80 db SPL and examined a range of attenuation for F_2, F_3 of from 10 to 40 db in 10-db steps. The experimental design allowed checks for internal consistency by providing in each stage at least one condition that was repeated in another stage. Each subject heard four tapes per session, each tape containing twelve blocks of nine trials (one monaural, binaural, and dichotic block for each signal/noise ratio, attenuation factor, or baseline SPL). The complete experiment occupied three sessions--one for each stage. The results have been plotted in Figure 1 and an analysis of the significance of the data is given in Table 1. The data plotted in Figure 1a show that a strong gain in listening performance emerges only when the baseline level is raised above 70 db SPL. At 80 db SPL the data points agree quite well with the data of Figure 1c

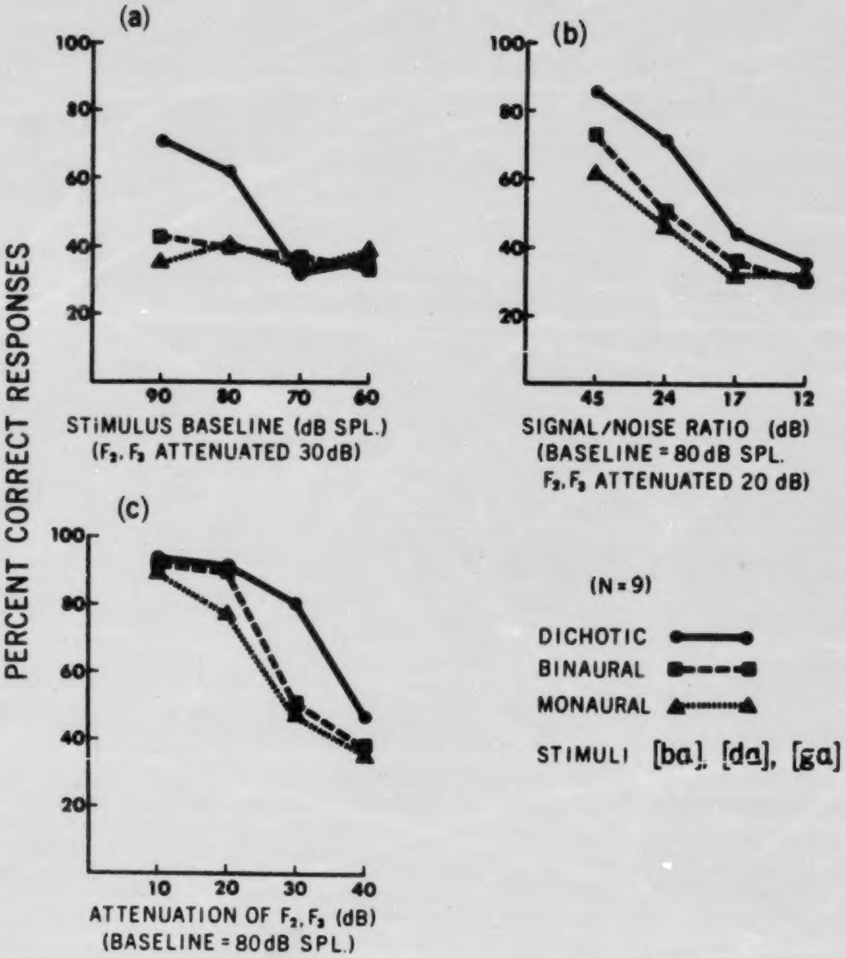

(a)
100
80
60
40
20
90
80
70
60
STIMULUS BASELINE (dB SPL.)
(F_2, F_3 ATTENUATED 30dB)
(b)
100
80
60
40
20
45
24
17
12
SIGNAL/NOISE RATIO (dB)
(BASELINE = 80dB SPL.
F_2, F_3 ATTENUATED 20 dB)
PERCENT CORRECT RESPONSES
(c)
100
80
60
40
20
10
20
30
40
ATTENUATION OF F_2, F_3 (dB)
(BASELINE = 80dB SPL.)
(N = 9)
DICHOTIC
BINAURAL
MONAURAL
STIMULI [ba], [da], [ga]

FIGURE 1

TABLE 1

Attenuation	Probability that Dichotic = Binaural
-10 db	$p > 0.8$
-20 db	$p > 0.6$
-30 db	$p < 0.001$
-40 db	$p \sim 0.008$
S/N Ratio	
45 db	$p \sim 0.008$
24 db	$p < 0.001$
17 db	$p \sim 0.01$
12 db	$p \sim 0.007$
Baseline	
90 db	$p < 0.001$
80 db	$p < 0.001$
70 db	$p > 0.02$
60 db	$p > 0.7$

obtained independently at an F_2, F_3 attenuation of 30 db. Figure 1c incidentally exhibits the "classic" differences among binaural, monaural, and dichotic listening performances in the face of F_2, F_3 attenuation.

The differences among the three listening conditions as a function of signal/noise ratio is shown in Figure 1b. At a signal/noise ratio of 45 db (essentially no noise) the dichotic mode of listening is superior to both binaural and monaural modes and the data points are in fair agreement with the corresponding independently measured points in Figure 1c. However, as the noise conditions become increasingly adverse, the performance in all three listening conditions is found to fall rapidly, reaching the chance response level at a signal/noise ratio in the region of 12 db.

Tests for a Right-Ear Advantage

Since the original reports by Kimura (1961a, 1961b) it has been known that dichotic competition among certain classes of speech sounds reveals a right-ear advantage. A number of recent studies (e.g., Day and Cutting, 1970; Darwin, 1971) have in particular examined the vocal features that appear to compete for the speech processing capability available in the left hemisphere of the brain. Because this study was concerned with a form of dichotic listening, it seemed reasonable to examine the data for an ear advantage which might show, for example, an enhancement of the release from masking for a particular ear/formant relationship. However, the results of Experiment 1 and four other experiments not reported here have revealed no ear advantage.

Experiment 2

Continuing the investigation of the effects of added noise, Experiment 2 compared monaural with dichotic listening performances for signals at a baseline level of 80 db SPL and a signal/noise ratio of 12 db. On this occasion the formants F_2 and F_3 were maintained at their natural levels relative to F_1 in order to find out whether the loss of dichotic superiority in Experiment 1 at a 12 db signal/noise ratio was due to the 20 db attenuation applied to F_2 and F_3.

In part one of the experiment the three voiced stop-vowel syllables [ba], [da], and [ga] were again used in randomized sequences recorded on three tapes. Each tape contained six blocks of twelve stimuli. Two blocks were dichotic, two were monaural, and the remaining two blocks contained only F_2 and F_3 presented monaurally. Noise was added to the stimuli in two parts to achieve an overall signal/noise ratio of +12 db with reference to the baseline signal of 80 db SPL. The first noise component was low-pass filtered and added to F_1 on channel A of the tape recorder and the second component was high-pass filtered and added to F_2, F_3 on channel B. The high/low crossover point (the 6 db down point) was set at 1 kHz, midway between F_1 and F_2. Monaural syllables were mixed with full band noise. Eleven subjects each heard four tapes in which ear/formant relations were balanced for both dichotic and monaural stimuli. The results are shown in Table 2.

TABLE 2

	Subject	Dichotic	Monaural	F_2, F_3 Alone
	1	68	65	63
	2	64	46	35
	3	79	58	58
	4	67	59	52
	5	64	64	61
	6	66	63	56
	7	74	71	61
	8	60	57	44
	9	59	58	52
	10	63	73	56
	11	67	57	54
Average		66.45	61.00	53.81

Value of "t"	Condition	Significance
2.105	Dichotic v Monaural	$p = 0.06$
4.604	Monaural v F_2, F_3 alone	$p > 0.001$
5.519	Dichotic v F_2, F_3 alone	$p > 0.001$

The second part of the experiment was a rerun of the first with the monaural blocks changed to binaural presentation. Six subjects listened to six tapes giving rise to 50 percent more data than was obtained in the first part of the experiment. The baseline signal intensity was again set at 80 db SPL. The results of this experiment are presented in Table 3.

Pooled results listed in Table 2 show that, although the dichotic performance was higher than the monaural performance, the calculation of "t" (on differences between performances on a subject-by-subject basis) indicates that the

TABLE 3

	Subject	Dichotic	Binaural	F_2, F_3 Bin
	1	108	99	112
	2	97	97	109
	3	93	93	96
	4	93	94	98
	5	95	95	90
	6	98	114	82
Average		97.33	98.67	97.83

Note: No differences were significant above $p = 0.1$ level.

overall difference is significant only at the $p = 0.06$ level. In part two, binaural performance was superior to dichotic by a small margin not statistically significant. The scores obtained on F_2 and F_3 alone in both experiments were high and again not significantly different from the scores obtained binaurally. This was not particularly surprising because the stimuli employed in the experiment demanded only place discriminations, the information carried in the upper pair of formants.

Experiment 3

High performances on F_2 and F_3 discriminated alone were also evident in another experiment. In this case the same stop-vowel syllables were employed in two dichotic tapes, one binaural tape, and one monaural tape in which only the F_2 and F_3 components were available. Attenuation of the F_2 and F_3 formants covered the range from 20 db to 50 db. The pooled data of all nine subjects employed in this experiment have been plotted in Figure 2.

The results show that the F_2 and F_3 stimuli heard alone fared even better than did the dichotic stimuli. This observation appears to be indicative of central masking, and the fact that the addition of a redundant F_1 actually degrades performance is certainly consistent with this interpretation whether or not the signals are being fused and interpreted as speech.

Experiment 4

Although the stop-vowel syllables used in the previous experiments are flexible stimuli for which the synthesis cues are well known, these syllables could not be relied upon to yield results applicable to the full repertoire of American English phones. Hence a stimulus set consisting of a variety of words was sought. The word list and procedure chosen for this experiment was the Modified Rhyme Test (MRT) developed by House, Williams, Hecker, and Kryter (1965) from an original formulation by Fairbanks (1958). Its closed response design makes the MRT easy to score on the basis of words correctly reported, although a full analysis of the phonetic confusions is often an involved procedure.

The MRT consists of 300 monosyllabic words grouped in blocks of six words which share the same vowel but differ in either their initial or final consonant.

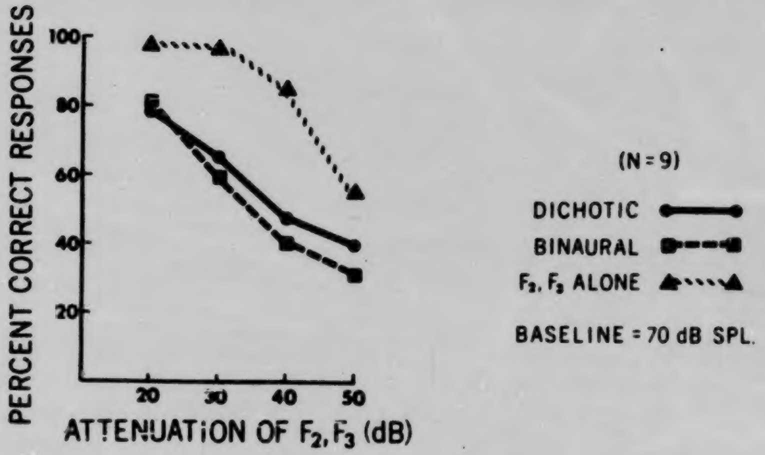

PERCENT CORRECT RESPONSES
100
80
60
40
20
20
30
40
50
ATTENUATION OF F_2, F_3 (dB)
(N=9)
DICHOTIC
BINAURAL
F_2, F_3 ALONE
BASELINE = 70 dB SPL.

FIGURE 2

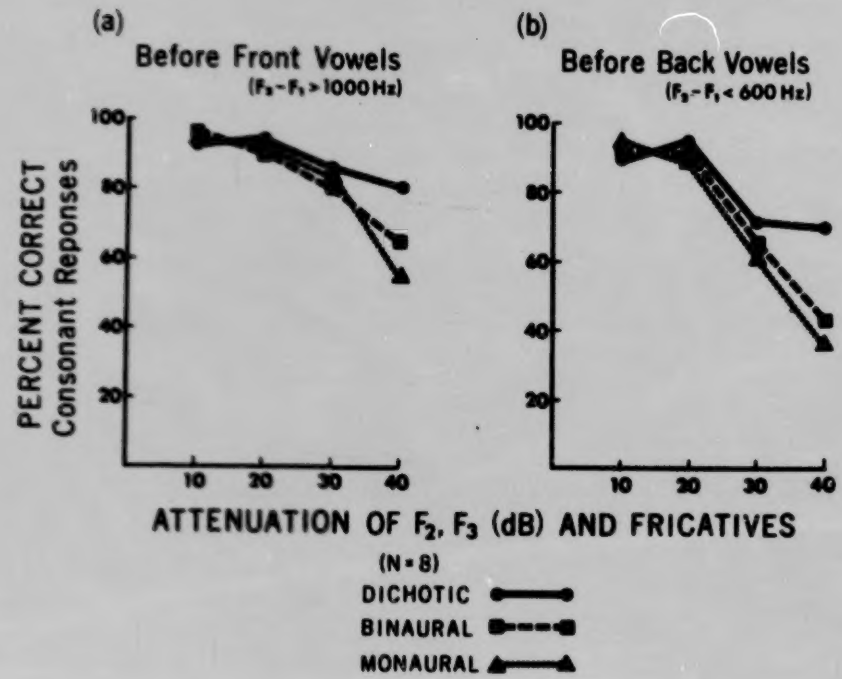
At each trial, one word from each block is presented in a carrier phrase of the form "Please mark the word" followed by a short pause and then the test word. The subject must mark one of the six words in each block which most closely resembles the word he heard. Two randomizations of the 300 words were recorded for monaural, binaural, and dichotic presentation. In all cases formant F_1 was recorded on channel B and formants F_2 and F_3 (in addition to the fricative signals) were recorded on channel A. Response sheets were printed with the word order within blocks--and the blocks themselves--randomized from session to session. The test words were presented at a baseline intensity of 80 db SPL while channel A (F_2 and F_3 and fricatives) was attenuated in 10-db steps over the range of 10 - 40 db. The binaural and monaural stimuli were produced by mixing the signals emerging from channels A and B of the recorder and directing the combined signals to one or both ears.

Eight listeners were employed. Their data have been plotted in Figures 3 and 4.

The results show that dichotic, monaural, and binaural listening conditions were indistinguishable from one another for combined F_2, F_3 and fricative attenuations of 10 and 20 db. For further increases in attenuation, however, monaural and binaural performances both fell more rapidly than was the case dichotically. Closer analysis illustrated in Figure 3 shows that initial consonant transitions into front vowels (which have F_2 loci) are less severely masked than transitions into the low back vowels. Moreover, this tendency dominates (as the earlier experiments predict) in conditions of high F_2, F_3 attenuation. In Figure 4 are plotted the recognition scores for the words (overall) and for stops, resonants, nasals, and fricatives as a function of combined F_2, F_3 and fricative attenuation. Resonants and nasals are heard particularly well under high attenuation in dichotic conditions, whereas the stops and fricatives, although also recognized more easily when heard dichotically, do not perform as well when F_2, F_3 and fricatives are attenuated. The rise in the dichotic recognition curve for nasals as the attenuation of F_2, F_3 increases should be interpreted with some caution. This is indicated by the fact that the synthesizer, which generated the stimuli, has no true nasal resonance and achieves a simulated nasality by manipulation of the intensities of oral formant filters. A second reason for caution is that the experiment employed naive listeners whose performance rise may come from learning. We will consider the consequences of learning effects in a later discussion.

Experiment 5

Experiment 5 represented an initial step in a study of the effects of dichotic F_1 versus F_2, F_3 listening on the intelligibility of continuous speech. Fifty nonsense sentences were constructed from monosyllabic words selected from the Thorndike and Lorge (1968) lists of the 200 most frequently used words in English. The sentences were all of the form "The (adjective) (noun) (verb, past tense) the (noun)" and are listed in Nye and Gaitenby (1974). Recordings of the sentences were made from the Haskins Laboratories' parallel formant resonance synthesizer in dichotic and binaural modes using the synthesis-by-rule algorithm described by Mattingly (1968). Formant F_1 was recorded on channel B and formants F_2 and F_3 together with all fricative signals were recorded on channel A. The pitch contour for the sentences descended in the manner characteristic of a statement, while the stress pattern took the form "The (mid) (high) (mid) the (high)"--with both utterances of "the" receiving low stress. The baseline signal level was set at 85 db SPL. Five subjects were paid to listen to the sentences



BASELINE = 80 dB SPL.
STIMULI FROM MODIFIED
RHYME TEST

FIGURE 3

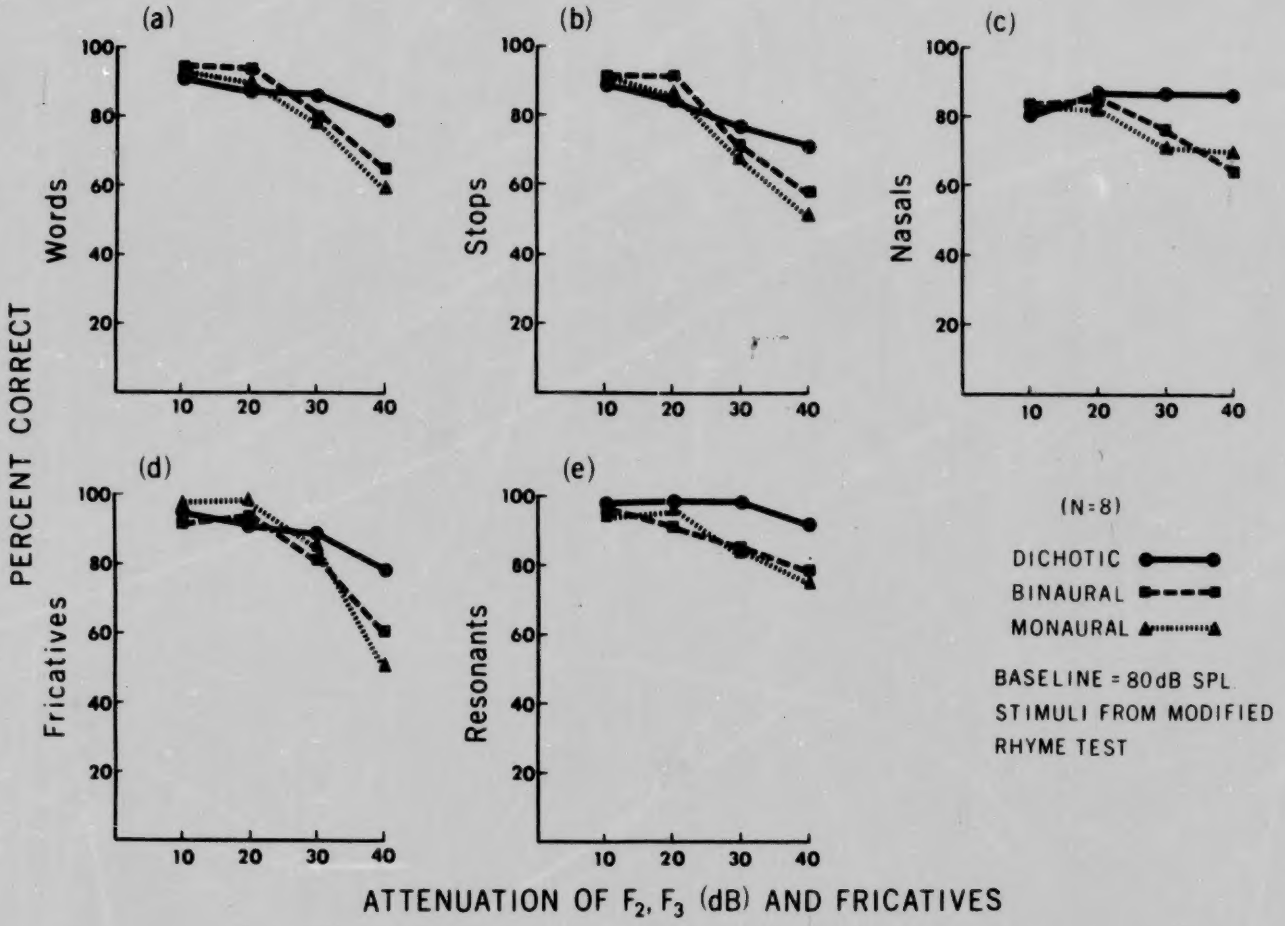

(a)
(b)
(c)
(d)
(e)
PERCENT CORRECT
Words
Stops
Nasals
Fricatives
Resonants
100
80
60
40
20
10
20
30
40
(N=8)
DICHOTIC
BINAURAL
MONAURAL
BASELINE = 80dB SPL
STIMULI FROM MODIFIED
RHYME TEST
ATTENUATION OF F2, F3 (dB) AND FRICATIVES

FIGURE 4

presented in blocks of five at F_2, F_3 and fricative attenuation levels of 0, 8, 16, 24, and 32 db. Binaural and dichotic modes of presentation were alternated in successive blocks and the experiment proceeded step by step in the order of increasing F_2, F_3 and fricative attenuation. Each subject wrote down the words he thought he heard and these data were subsequently analyzed to derive percentages of the words correctly identified.

Figure 5 contains a graph of the pooled data that were obtained in order of increasing F_2, F_3 and fricative attenuation from Experiment 5. The performance is seen to rise to a peak at an attenuation of 16 db and to fall rapidly at further levels of attenuation. The reason for the performance maximum is indicated by a considerable body of data now being collected on the process of learning synthetic speech. Performance on the interpretation of synthetic speech does improve with prolonged exposure. Thus, the curve of Figure 5 is the result of a rapid rise in the subjects' familiarity with the special characteristics of synthetic speech, superimposed upon a steady decline in intelligibility of the speech with increasing attenuation. However, despite the distortion imposed by learning effects, the superiority of dichotic listening at high levels of combined F_2, F_3 and fricative attenuation clearly emerges.

DISCUSSION

Relevance to Natural Speech Perception

As experimental work has proceeded, several technical difficulties and potential pitfalls have become apparent in two specific areas.

First, there are many stimulus dimensions of interest in the dichotic release from masking phenomenon, and to embrace in detail a wide range of signal/noise ratios, baseline intensity levels, and consonant-vowel pairs would be a formidable task. However, without such an extensive analysis, it will probably not be possible to give a completely accurate assessment of the full benefit (if any) to be gained from dichotic listening. The results of this study can only point out promising directions, and firm conclusions can be drawn only for the conditions which have been actually examined.

Second, while we feel confident of the accuracy of the data we have gathered using synthesis procedures, there is a legitimate question about the extent to which these results have a bearing on the perception of real speech. From the fact that listeners recognize the output of a synthesizer as speech, however, one must conclude that the data are, to a degree, significant. What appears to be at issue is whether, in natural speech, masking phenomena are of lesser importance or possibly of more importance than they are in synthetic speech. This question must eventually be confronted and efforts are now under way to explore listeners' performances with real-speech sounds.

Learning Effects

Yet another topic of concern is the matter of listening experience. As the repertoire of speech sounds being examined is enlarged, so learning effects become more evident and tend to disturb and confuse the data. Experienced subjects are difficult to obtain, and this fact emphasizes the need to utilize natural speech sounds that will be familiar and therefore place less pressure on the subject to make special perceptual allowances. However, subjects will still be required

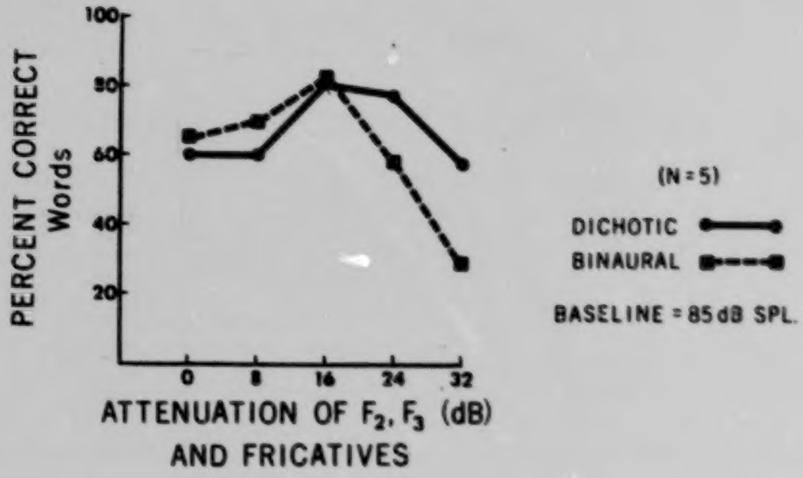

PERCENT CORRECT
Words
100
80
60
40
20
0
8
16
24
32
ATTENUATION OF F_2, F_3 (dB)
AND FRICATIVES
(N = 5)
DICHOTIC
BINAURAL
BASELINE = 85 dB SPL.

FIGURE 5

to learn to accommodate the dichotic mode of listening and it is evident that further effort will have to be invested in exploring the question of fusion.

Selecting a Suitable Noise Spectrum

Further problems have emerged in the selection of appropriately shaped noise spectra intended to create listening conditions closely paralleling those that might be engineered in a real-speech communications system. The choice actually adopted in the experiments involved the division of the noise spectrum at 1000 Hz; however, while such a dividing point satisfactorily bisects F_1 and F_2 for the vowel /ɑ/ it does not meet the requirements of other back vowels nor in particular the stop transitions into those vowels which frequently cross the 1000 Hz boundary. Thus an F_2 transition can originate in a noise-free region of the spectrum and cross the boundary into the F_2 locus for the vowel. The problem is that its brief exposure free of noise makes a transition more easily identifiable than would otherwise be the case. On balance, however, the present method of allocating noise to the F_1 and F_2, F_3 channels appears to be the best, but it points up once again the fact that the results obtained in this manner may not extrapolate well into a real-life situation.

Fusion of Dichotic Stimuli

The fact that to a large extent F_1 is redundant in voiced stop-plus-vowel syllables where the vowel is always the same means that the listener does not need to fuse the dichotic stimuli to arrive at a decision. Moreover, because fusion is essential to the interpretation of dichotic stimuli as speech, it is obviously necessary to examine the fusion issue more closely and to extend the analytical tests toward a wider variety of speech sounds. Several factors, however, point to the conclusion that fusion does indeed occur.

First are the subjective impressions of the experimenters who, although being aware of the stimulus composition, do not consciously find themselves paying special attention to one ear or to the other when listening to dichotic stimuli. To these observers, stimuli with high F_2, F_3 attenuation appear to be entirely monaural but the fact that the scores that can be achieved are higher in this condition than when the attenuated F_2 and F_3 are combined binaurally with F_1 indicates the entirely unconscious nature of dichotic listening and the process of fusion. Furthermore, there is a growing body of evidence in the literature that fusion can and does take place with speech stimuli. Broadbent and Ladefoged (1957) discuss the question of the fusion of separated formants in broad terms, showing that when exposed to a two-formant dichotic mode of presentation, a majority of subjects will report hearing only one voice in one location. More relevant to the question of fusion (as we are concerned with it here) is an experiment briefly alluded to by Carlson, Granström, and Fant (1970) in connection with an investigation that dealt largely with the so-called "second spectral peak of front vowels" (Fujimura, 1967). In the course of a number of experiments on isolated vowels, Carlson et al. describe a dichotic presentation of vowel sounds in the following words:

> "...by connecting the outputs of the formant circuits to two different channels. Thus the different channels could supply different ears or both ears could be exposed to the whole signal. The vowel identity proved to be invariant to such changes, leaving space impressions and minor timbre changes as distinguishing cues" p. 32.

Again the evidence suggests that fusion does occur. Work is currently in progress on experiments that are expected to reveal some characteristics of the subjective phenomenon of fusion. The results will be reported in a future paper.

REFERENCES

Broadbent, D. E. and P. Ladefoged. (1957) On the fusion of sounds reaching different sense organs. J. Acoust. Soc. Amer. 29, 708-710.

Carlson, R., B. Granström, and G. Fant. (1970) Some studies concerning perception of isolated vowels. Quarterly Progress Status Report (Speech Technology Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR 2-3, 19-35.

Darwin, C. J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 46-62.

Day, R. S. and J. E. Cutting. (1970) Perceptual competition between speech and nonspeech. J. Acoust. Soc. Amer. 49, 85(A). (Also in Haskins Laboratories Status Report on Speech Research SR-24, 35-46.)

Fairbanks, G. (1958) Test of phonemic differentiation: The Rhyme Test. J. Acoust. Soc. Amer. 30, 596-600.

Fant, G. M. (1950) On the predictability of formant levels and spectrum envelopes from formant frequencies. In *For Roman Jacobson*, ed. by M. Halle, H. Lunt, and H. MacLean. (The Hague: Mouton).

Fujimura, O. (1967) On the second spectral peak of front vowels: A perceptual study of the role of the second and third formants. Lang. Speech 10, 181-193.

House, A. S., C. E. Williams, M. H. L. Hecker, and K. D. Kryter. (1965) Articulation-testing methods: Consonantal differentiation with a closed-response set. J. Acoust. Soc. Amer. 37, 158-166.

Kimura, D. (1961a) Some effects of temporal-lobe damage on auditory perception. Canad. J. Psychol. 15, 156-165.

Kimura, D. (1961b) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Mattingly, I. G. (1968) Synthesis by rule of General American English. Ph.D. dissertation, Yale University. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Nye, P. W. and J. G. Gaitenby. (1974) The intelligibility of synthetic speech in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research SR-37/38 (this issue).

Rand, T. C. (1974) Dichotic release from masking for speech. J. Acoust. Soc. Amer. 55, 678-680. [Also in Haskins Laboratories Status Report on Speech Research SR-33, 47-55 (1973).]

Thorndike, E. L. and I. Lorge. (1968) *The Teacher's Word Book of 30,000 Words*. (New York: Teacher's College Press).

Binaural Subjective Tones and Melodies Without Monaural Familiarity Cues*

Michael Kubovy,[+] James E. Cutting,[++] and Roderick McI. McGuire[++]

Julesz has shown that cross-correlations between two patterns that appear random to either eye alone can give rise to the perception of form and depth when viewed stereoscopically. We produced auditory analogs by presenting eight simultaneous and continuous sine waves to both ears and by either phase-shifting or frequency-shifting one sine wave relative to its counterpart in the opposite ear. Particular tones were shifted in sequence so that a melody was heard--a melody that was undetectable by either ear alone.

Julesz (1971) has shown that if one presents a field of random dots to one eye and the same field to the other eye, but with a small portion shifted horizontally, a certain area of the percept appears to stand out in depth. Its contour is the boundary of the shifted portion of dots, and the shift is logically impossible to detect by one eye alone. Julesz named this phenomenon _cyclopean_ perception after the mythical giants who looked out at the world through a single eye in mid-forehead. With random-dot stereograms or anaglyphs it is possible to bypass, as it were, the peripheral visual apparatus and project information to the cyclopean eye and onto "the 'mind's retina'--that is, at a place where the left and right visual pathways combine in the visual cortex" (Julesz, 1971:3). Our goal was to devise an auditory analog to the cyclopean percept, one for etymological reasons we call _cyclotean_,[1] in which the peripheral auditory apparatus is bypassed and information is projected onto the "mind's cochlea."

We were provoked into seeking this goal, in part, by Julesz (1971:51) and Julesz and Hirsh (1973) who claim that analogies between visual and auditory perception are not "very deep." The basis for their view is that visual perception

---

*Presented at the 87th meeting of the Acoustical Society of America, New York, April 1974.

[+] Department of Psychology, Yale University, New Haven, Conn.

[++] Haskins Laboratories and Yale University, New Haven, Conn.

[1] This neologism is constructed from Greek roots to be analogous to the term _cyclopean_. Since _ops_ is the Greek root for eye and _oto_ the root for ear, we feel that _cyclotean_ is the proper term for this phenomenon. As yet we have been unable to find reference to such a mythical being. One reason for this may be that the _cyclot_ was a much less truculent, and hence much less memorable, creature.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

is primarily concerned with spatial objects whereas auditory perception is primarily concerned with temporal events. The distinction between objects and events appears to be primarily founded on the potential richness of percept in each modality: two spatial dimensions are possible for a visual percept, whereas only the one temporal dimension is available for an auditory percept.

In precyclopean days, Huggins (reported by Licklider, 1956) and later Cramer and Huggins (1958) (and Fourcin, 1962; and Guttman, 1962) demonstrated that if one presents white noise to one ear and the same white noise to the other, but with a narrow band of frequencies time-delayed, a faint pitch quality is heard. It sounds like narrow-band filtered noise, and is logically impossible to detect with one ear alone. Nevertheless, it fails to meet Julesz' (1971:51) criterion for an auditory object: it is "not a truly cyclopean phenomenon since the input variable is a single time delay, while the perceived variable is a single pitch." Julesz elaborates by stating that to create a true analog to the visual phenomenon with such pitches one would need to generate a melody. Since a melody is a pattern of pitches and varies in both time and frequency, it is multidimensional and hence an auditory object. Thus, we decided to produce a cyclotean melody. In order to probe the generality of auditory analogs we chose to generate the melody by two conceptually distinct methods. The first is a methodological offshoot of Cramer and Huggins (1958) and is analogous to existing visual work; the second method, on the other hand, is wholly new.

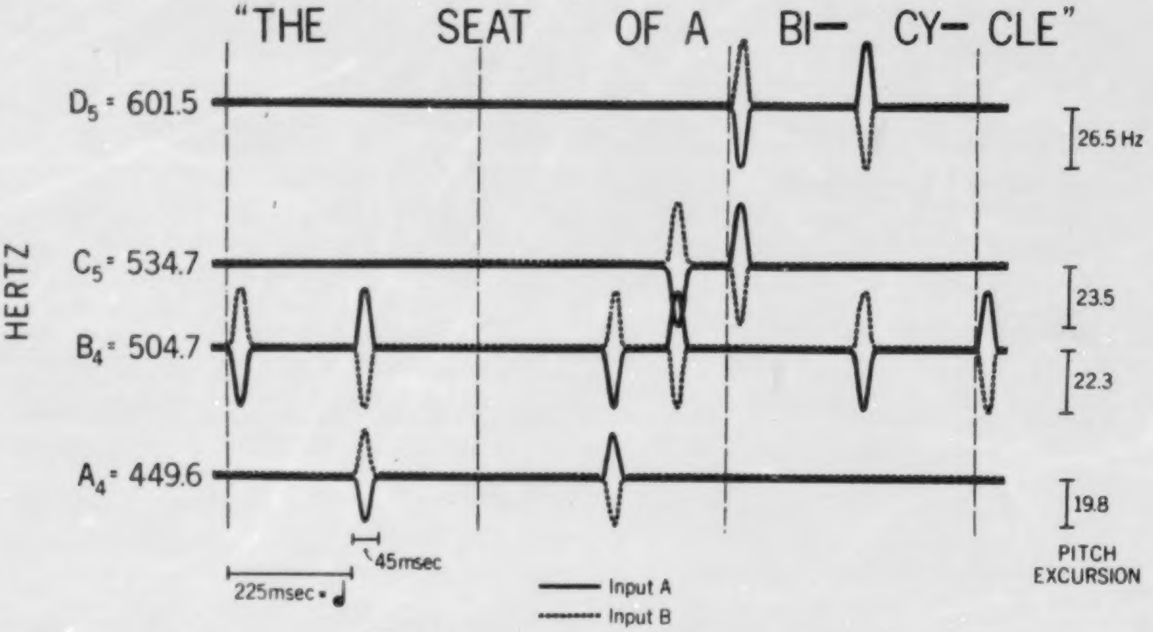
In both demonstrations the basic stimulus consists of eight simultaneous, continuous, computer-generated sine waves whose frequencies were chosen from the even-tempered scale in the key of G.[2] In both, the tune Daisy was embedded in the tonal arrays as a cyclotean melody. All stimuli were computed numerically, output by the Haskins Laboratories' PCM system (Cooper and Mattingly, 1969), and recorded at the same time on two channels of audio tape.

Our first demonstration begins with the presentation of the basic stimulus to both ears, with a lag of one msec between the onsets of the two inputs (Input A leading Input B). A constant discord of eight tones is heard for 1500 msec. At that point the first note, D_5, is introduced by advancing the phase of the D_5-component of Input B by one msec and by delaying the phase of its counterpart in Input A by the same amount. The phase-shifting process is not instantaneous, but occurs over a 45 msec duration. The first note is sustained by maintaining the new phase relation until 900 msec has elapsed after the initiation of the phase shift, at which time the phase-shifting process is reversed (again taking 45 msec) until the two corresponding sine waves resume their original phase relationship. The offset phase-shifting of D_5 completely overlaps with the onset of the subsequent note, B_4. Subsequent notes in the tune are introduced and removed in the same fashion. The duration of each note is between 112.5 msec for an eighth note and 1800 msec for a double-whole note. The duration of the entire sequence is approximately 24 sec. A spectral segment of it is represented in the top panel of Figure 1.

---

[2] The lowest note, D_4, was 300 Hz, whereas the highest note, D_5, was adjusted to 601.5 Hz to avoid harmonic relationships with D_4. The other six tones were left at even-tempered frequencies since their pairwise frequency ratios are all irrational.

a cyclotean melody from phase shifts



a cyclotean melody from frequency shifts

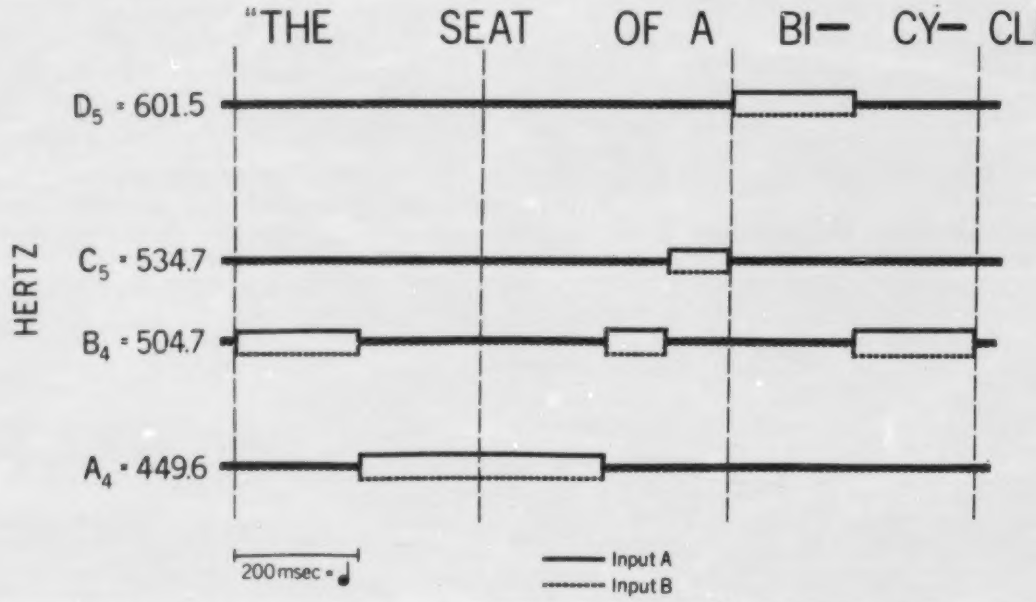


Figure 1: Schematic spectrograms of the four highest-frequency sine waves in the basic stimuli during that portion of Daisy corresponding to the lyric "upon the seat of a bicycle built for two." The frequencies of opposite-ear stimuli are superimposed. The top panel shows that in the first demonstration phase-shifting of opposite-ear stimuli was realized in terms of frequency changes. Since it is not possible to effect a change in phase without also temporarily changing pitch, pitch excursions were necessary. They were realized in raised and lowered cosine functions with maximal increase and decrease of 4.4 percent. The lower panel shows how frequency shifts were realized in the second demonstration for the same segment.

Subjectively, the melody is perceived to occur inside the head but displaced to one side of the midline, while a background noise is localized to the opposite side.[3] The notes are perceived to have a sparkling onset, much like that of a chime. Offsets, however, are not particularly striking in any way. When Input A or Input B is listened to alone, only occasional faint perturbations are audible in an otherwise continuous, noise-like signal.

All these observations were confirmed by 24 subjects in two experimental conditions. In the dichotic condition, where Input A was presented to one ear and Input B to the other, all subjects reported hearing a clear sequence of tones embedded in noise. Only a few subjects failed to recognize the tune.[4] On the other hand, in the diotic condition, where one input was presented alone to both ears, no subject reported hearing individual tones, let alone the melody.

The analogies between the preceding demonstration and Julesz' stereograms are fairly straightforward: interaural phase differences correspond to interocular disparities, auditory localization corresponds to visual depth perception, and melodic contour corresponds to visual form.

In the second demonstration, however, the analogies are less obvious: this cyclotean Daisy is generated by binaural beats (for discussion of binaural beats, see Licklider, Webster, and Hedlun, 1950; Perrot and Nelson, 1969; and Oster, 1973). Here the demonstration begins with the presentation of the basic stimulus to both ears, but with Input A 180 degrees out of phase with Input B. A discord of eight tones is heard for 1200 msec. Then D_5 is introduced by increasing the frequency of the D_5-component in Input A by 5 Hz and decreasing the frequency of its counterpart in Input B by the same amount. Frequency shifts are instantaneous (within a range of ±250 microsec accuracy). The first note is perceived to have the pitch of the original D_5 around which binaural beats of 10 Hz co-occur. D_5 is sustained for 800 msec, at which point the frequency shift is instantaneously reversed and the two corresponding sine waves resume their original phase and frequency relationships. The second note, B_4, is introduced by the same process at the offset of D_5, and all subsequent notes follow the same pattern. Each note is between 100 and 1600 msec in duration, thus embracing 1 to 16 beats. The duration of this second demonstration is approximately 22 sec. A spectral segment of it is represented in the lower panel of Figure 1.

Subjectively, the melody is again perceived to occur inside the head. It is not perceived to be localized differently from the background, and yet it does stand out as figure against ground. Beats are not always perceptible, and when they are, they do not appear to be part of the figure or ground but rather a disembodied roughness. Aside from the beats and localization, the new Daisy sounds similar to the first, but perceptually more prominent. Again, Inputs A and B are not sufficient by themselves to yield the percept. Occasional beats are audible in a single stimulus but these are due to monaural beats where adjacent sine waves are frequency-shifted to within 15 Hz of one another.

---

[3] With phase shifts of exactly one msec for each of the eight sine waves, the degree to which opposite-ear sine waves are out of phase is not optimal, but varies according to wavelength. The amount of interaural phase differences at different frequencies should therefore result in different perceived spatial locations. Critical listening can give rise to a spatially varying cyclotean percept, but few listeners can detect it without being instructed what to listen for and without having previously listened to the second demonstration.

[4] One subject, for example, was a New Zealander who reported hearing Waltzing Matilda, not Daisy. We attribute this misperception to his ethnocentricity and general musical inability, not to the subjective strength of the percept.

These observations were confirmed by a group of 12 subjects. All subjects heard and identified the melody dichotically, whereas none heard any notes diotically. After listening to the first and second demonstrations, most subjects agreed that the second rendition was perceptually more compelling.

Although from the point of view of auditory theory the first demonstration is intimately related to the domain of masking level difference (MLD), it deals with a qualitatively different phenomenon. Whereas MLD is investigated with stimuli that are barely detectable monaurally and better detected binaurally, this cyclotean effect deals with the binaural segregation of auditory figure from ground, where nothing can be perceived monaurally. We do not know, however, what the relation is between MLDs and our second demonstration, since MLDs have not been studied with relatively fast-beating stimuli and it is not known whether they can produce unmasking. Egan (1965) has used slowly beating stimuli to demonstrate MLDs, exploiting the principle that stimuli which differ slightly in frequency can be conceived of as stimuli with identical frequencies but with constantly changing phase relations. In his demonstration a signal is heard to fade in and out of noise with the beat frequency. No such fading occurs in our second demonstration. In addition, it should be noted that no visual analog of our second demonstration has yet been generated.[5] Further studies of the domain of cyclotean perception are presently under way.

REFERENCES

Blakemore, C. (1970) A new kind of stereoscopic vision. Vision Res. 10, 1181-1199.

Cooper, F. S. and I. G. Mattingly. (1969) Computer-controlled PCM system for investigation of dichotic speech perception. J. Acoust. Soc. Amer. 46, 115(A).

Cramer, E. M. and W. H. Huggins. (1958) Creation of pitch through binaural interaction. J. Acoust. Soc. Amer. 30, 413.

Egan, J. P. (1965) Demonstration of masking-level differences by binaural beats. J. Acoust. Soc. Amer. 37, 1143-1144.

Fourcin, A. J. (1962) An aspect of the perception of pitch. In Proceedings of the Fourth International Congress of Phonetic Sciences. (The Hague: Mouton) 355-359.

Guttman, N. (1962) On defining the range of pitch perception. J. Acoust. Soc. Amer. 33, 862(A).

Julesz, Bela. (1971) Foundations of the Cyclopean Perception. (Chicago, Ill.: University of Chicago Press).

---

[5] Any attempt to generate such a pattern would probably rely on previous work by Blakemore (1970), who used binocular combination of gratings.

Julesz, Bela and I. J. Hirsh. (1973) Visual and auditory perception: An essay of comparison. In Human Communication: A Unified View, ed. by E. E. David and P. Denes. (New York: McGraw-Hill) 283-340.

Licklider, J. C. R. (1956) Auditory frequency analysis. In Information Theory, Third London Symposium, ed. by C. Cherry. (London: Butterworth) 253-368.

Licklider, J. C. R., J. C. Webster, and J. M. Hedlun. (1950) On the frequency limits of binaural beats. J. Acoust. Soc. Amer. 22, 468-473.

Oster, G. (1973) Auditory beats in the brain. Sci. Amer. 229, 94-103.

Perrot, D. R. and M. A. Nelson. (1969) Limits for the detection of binaural beats. J. Acoust. Soc. Amer. 46, 1477-1481.

Categories and Boundaries in Speech and Music

James E. Cutting[+] and Burton S. Rosner[++]

Perceptual categories and boundaries arise when subjects respond to continuous variation on a physical dimension in a discontinuous fashion. It is harder to discriminate between members of the same category than to discriminate between members of different categories, even though the amount of physical difference between both pairs is the same. Speech stimuli have been the sole class of auditory signals to yield such perception; for example, each different consonant phoneme serves as a category label. Experiment I demonstrates that categories and boundaries occur for both speech _and_ nonspeech stimuli differing in rise time. Experiment II shows that rise time cues categorical differences in both complex and simple nonspeech waveforms. Taken together, these results suggest that certain aspects of speech perception are intimately related to processes and mechanisms exploited in other domains. The many categories in speech may be based on categories that occur elsewhere in auditory perception.

Speech is replete with perceptual categories, as so much research at the Haskins Laboratories has shown (see Liberman, Harris, Hoffman, and Griffith, 1957; Lisker and Abramson, 1964; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Pisoni, 1971, 1973; among many others). Different phonemes, particularly the stop consonants, serve as category labels when synthetic speech stimuli are varied along a particular acoustic dimension and the listener must identify each item on the array. Typically, continuous physical change yields markedly discontinuous percepts. For example, when the slope of the second-formant transition in certain speech patterns is varied in equal steps by adjusting its starting frequency, the resulting stimuli are perceived as [ba] or [da] or [ga] but never as anything else and rarely as anything in between. Not only are these identifications quantal, but listeners often discriminate poorly between two acoustically different but phonetically identical items. In contrast, subjects can more easily discriminate items separated by the same amount of acoustic difference when accompanied by a phonetic difference. This curious nonlinearity is known as categorical perception. The discontinuity in the discrimination function points to the locus of a perceptual boundary.

Auditory signals carrying no linguistic information (so-called "nonspeech" sounds) might also have categories and boundaries. Clearly, however, many

---

[+] Yale University and Haskins Laboratories, New Haven, Conn.

[++] University of Pennsylvania, Philadelphia.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

nonspeech sounds do not. No stable boundaries occur for sine waves of different frequencies (Sawusch and Pisoni, 1974). None have been found in ABX discrimination tasks for spectral inversions of speech stimuli (Liberman, Harris, Kinney, and Lane, 1961); for "chirps" and "bleats," which are brief segments of speech stimuli that carry important phonetic information in a speech context (Mattingly, Liberman, Syrdal, and Halwes, 1971); or for speech-like sounds with phonetically irrelevant formant transitions (Cutting, in press). Should we conclude that nonspeech sounds are never perceived categorically? We think not. The usual explorations of categorical perception in nonspeech realms have used stimuli that are either too simple or too heavily tied to speech patterns (single formants, single formant transitions, and clusters of distorted formants). Nonspeech sounds which occur in our everyday environment might have categories and boundaries, and might be perceived categorically like speech stimuli.

A logical candidate here is musical sounds. Thus, the present studies investigated the identification and discrimination of selected music-like sounds and, for comparison, of speech syllables. The dimension we chose to vary is rapidity of stimulus onset, called attack or rise time.

EXPERIMENT I

Method

Stimuli. Two classes of stimuli were synthesized for identification and discrimination: 18 speech stimuli and 18 nonspeech stimuli. Nonspeech stimuli were sawtooth waves generated on the Moog synthesizer at the Presser Electronic Studio at the University of Pennsylvania. Two nine-item arrays consisted of stimuli that differed solely in their onset characteristics. One array was synthesized at 440 Hz, and the other at 294 Hz. Amplitude envelopes reached maximum intensity in 0, 10, 20, 30, 40, 50, 60, 70, or 80 msec after onset. By 0 rise time we mean that a stimulus reached maximum amplitude in one-fourth of a period. Rise times were measured by digitizing the waveforms and displaying them with high resolution on a computer-controlled oscilloscope. The rapid-onset stimuli sounded like the plucking of a stringed instrument whereas the slower-onset stimuli sounded like the playing of the same instrument with a bow. The durations of the nonspeech stimuli were between 1020 and 1100 msec, varying according to rise time. Oscillograms of stimuli with 10 and 70 msec rise times are shown in Figure 1. The sawtooth stimuli had some low-frequency amplitude modulation, due to the Moog oscillator.

Speech stimuli were generated on the Haskins Laboratories' parallel resonance synthesizer. Like the nonspeech stimuli, they formed two nine-item arrays, with members of each array differing in rise time by 10-msec increments, from 0 to 80 msec. In one array items were identifiable as either [tʃa] as in CHOP or [ʃa] as in SHOP, and in the other array as either [tʃæ] as in CHAD or [ʃæ] as in SHAD. All speech stimuli shared the same pitch contour and were between 410 and 490 msec in duration, differing again according to rise time. Oscillograms of speech syllables with 10 and 70 msec rise times appear in Figure 1. Sine waves with 10 and 70 msec rise times are shown for comparison.

All stimuli were recorded on audio tape, then digitized, edited, and stored on a disc file using the PCM system at Haskins Laboratories (Cooper and Mattingly, 1969). Stimuli were reconverted to analog form at the time test tapes were recorded for each task.

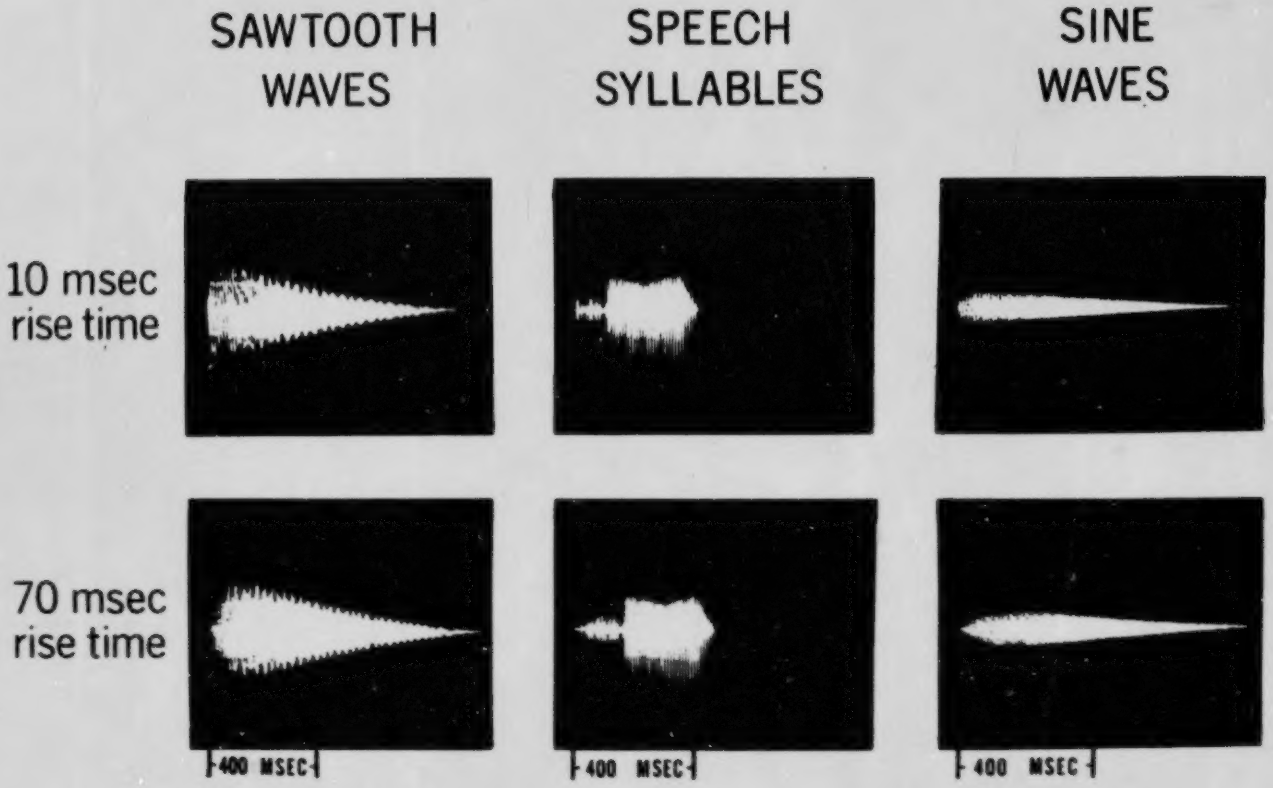


FIGURE 1

Figure 1: Oscillograms of sample stimuli used in Experiments I and II.

Tapes. Two identification and two discrimination tapes were recorded, one of each for the two classes of stimuli. Identification tapes consisted of a random sequence of 144 items: (2 arrays) X (9 items per array) X (8 observations per item). Onset-to-onset time was 4 sec with a 7-sec pause between blocks of 12 items.

Discrimination tapes consisted of ABX stimulus triads with 2 sec between onsets of items within a triad and 5 sec between triads. All four permutations of each comparison were represented: ABA, ABB, BAB, and BAA. Stimulus A and Stimulus B were members of the same array and differed in rise time by 20 msec. Thus, there were seven possible comparisons of rise times (in msec) for each array: 0 and 20, 10 and 30, 20 and 40, 30 and 50, 40 and 60, 50 and 70, and 60 and 80. Discrimination tapes consisted of a random sequence of 56 triads: (2 arrays per stimulus class) X (7 comparisons per array) X (4 ABX permutations).

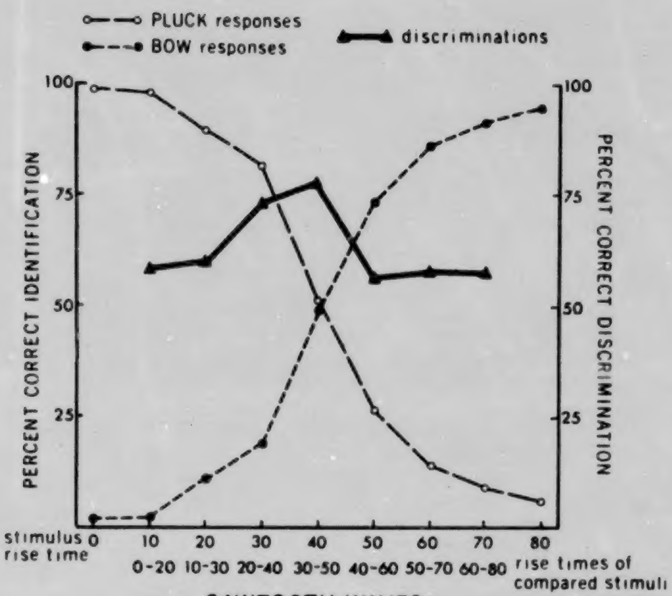
Subjects, apparatus, and procedure. Twenty Yale University undergraduate students participated in two tasks on each of two testing days as part of a course requirement. Audio tapes were played on an Ampex AG-500 tape recorder and broadcast over an Ampex 620 loudspeaker in a partially sound-attenuating room. On the first day subjects listened to the nonspeech sounds. For preliminary training, the endpoint stimuli were played in an alternating sequence five times each at both frequencies. Subjects were told to regard the 0 msec stimuli as a *plucked* string of a musical instrument (like that of a guitar) and to regard the 80 msec stimuli as a *bowed* string of a musical instrument (like that of a violin). Subjects readily agreed that these labels were easy to use. During the identification task they checked off their response to each stimulus, *pluck* or *bow*, on a prepared response sheet. During the discrimination task they listened to each triad and wrote *A* or *B*, indicating which of the first two items in the triad they felt was identical to the third item.

On the second day subjects listened to the speech sounds. The tasks and basic instructions were the same as the first day, except that during the identification task subjects checked off their response, *CH* or *SH*, for each item.

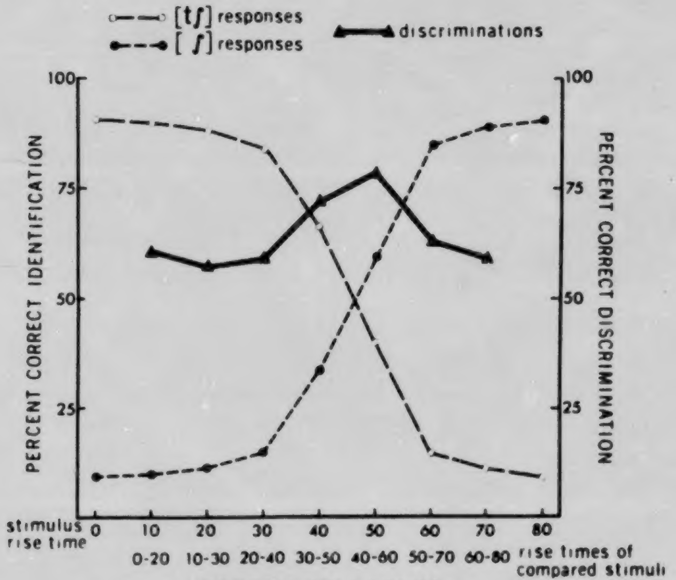
Results and Discussion

Sawtooth waves. The top panel of Figure 2 shows that identifications of the nonspeech stimuli were quite categorical. It combines results for the 294 and 440 Hz arrays, which did not differ in their effects. The stimuli with rise times of 0, 10, 20, and 30 msec were identified as a plucked sound on 92 percent of all trials; the 50, 60, 70, and 80 msec stimuli were identified as bowed sounds on 87 percent of all trials; and only the 40-msec stimulus was ambiguous. The ABX discrimination function is overlaid on the identification results and shows a pronounced peak at the rise-time comparisons of 20-40 msec and 30-50 msec. Each point is significantly greater by a sign test than its adjacent within-category comparison, 10-30 msec ($z = 2.23$, $p < .02$) and 40-60 msec ($z = 3.65$, $p < .001$). There were no significant differences among the other discriminations. We checked whether these results reflect clicks induced at short rise times in the loudspeaker, and found such artifacts only at 0 and 10 msec rise times but not at 20 msec and longer.

Speech syllables. The identification and discrimination functions for all of the speech stimuli appear in the lower panel of Figure 2. Stimuli with rise times of 0, 10, 20, and 30 msec were identified as beginning with [tʃ] on 88



SAWTOOTH WAVES



SPEECH SYLLABLES

Figure 2: Identification and discrimination functions for sawtooth waves and speech syllables.

percent of all trials; stimuli with 60, 70, and 80 msec rise times were identified as beginning with [ʃ] on 80 percent of all trials; and the 40 and 50 msec stimuli were ambiguous. The discrimination function peaks at the rise-time comparisons of 30-50 msec and 40-60 msec. Each point is significantly greater than its adjacent within-category comparison, 20-40 msec ($\underline{z} = 2.36$, $\underline{p} < .01$) and 50-70 msec ($\underline{z} = 2.56$, $\underline{p} < .01$). There were no significant differences among the other comparisons, nor, of course, between the [tʃa]-[ʃa] and the [tʃæ]-[ʃæ] arrays.

A comparison of speech and music. The results with the two classes of stimuli are remarkably parallel: speech and musical sounds both yielded reasonably quantal identification functions and moderately peaked discrimination functions. That the results are not more impressive probably reflects two causes: the stimuli were played over a speaker rather than through earphones, and the points displayed in Figure 2 are group data. Individual data in this experiment tend to show higher discrimination peaks and sharper category boundaries; averaging softens these extremes. For a comparison with other speech results in identification and discrimination, see Pisoni (1971). He also discusses the relative merits of the ABX task compared with other discrimination tasks.

Category boundaries for the two classes of stimuli were at somewhat different points, between 30 and 40 msec for the nonspeech and between 40 and 50 for the speech stimuli, but this does not impair the overall similarity of the results. The similarity, in fact, is considerably greater than might have been expected. The stimulus classes differ radically in the aspect of the signal which has been varied: the nonspeech stimuli are periodic throughout, whereas the speech stimuli are aperiodic during the portion of the stimulus that contains the variation. Furthermore, rise time is essentially the only variable that separates the nonspeech items. In the ABX task, stimuli within a triad differed in duration by only 20 msec. This is far below the difference limen for sounds of approximately 1 sec duration (Fraisse, 1963). In contrast, rise time and duration are cues for the affricate/fricative segment of the speech syllables. In this case, the durational differences equal or exceed the difference limen. Gerstman (1957) noted that either rise time or duration can cue this distinction, but neither cue by itself is as potent as the two together. Thus, in some sense, the choice of stimuli in the present experiment differentially favored categorical perception of the speech stimuli in that two important acoustic cues were varied instead of one. The results, however, do not reflect this advantage.

We now must consider why plucked and bowed notes "behave like speech;" that is, why sawtooth waves demonstrate categories and boundaries. One possibility is that the nonspeech discrimination function shown in Figure 2 occurred because of the verbal labels pluck and bow. Since the identification task preceded the discrimination task, subjects had as much as 20 minutes of practice at labeling the sounds. Perhaps the effects of this practice carried over into the ABX task, and discriminations were mediated by labels. A likely way to eliminate this flaw is to reverse the order of the tasks and have subjects perform the ABX task before the identifications. This was done in Experiment II.

Categories may occur for sawtooth stimuli because their waveforms have complex spectra or because they sound similar to stringed instruments being played, a familiar occurrence in our environment. Categorical perception may not occur for simpler sounds. In other words, does this boundary between pluck and bow occur for all sounds that vary in rise time, or just for certain sounds? Experiment II also addressed this question.

EXPERIMENT II

Method

The sawtooth identification and discrimination tapes used in Experiment I were employed here as well. Another set of 18 stimuli was generated: it consisted of sine waves varied in exactly the same manner as the sawtooth stimuli (at both 294 and 440 Hz). They were arranged in the same random orders as the sawtooth series for the two tasks, and corresponding identification and discrimination tapes were recorded. Sine-wave stimuli with 10 and 70 msec rise times are shown in Figure 1. In an effort to sharpen up the identification and discrimination functions, signals were presented through matched Telephonics earphones (Model TDH-39). Twelve Yale University undergraduate students, who served as paid volunteers, listened to both discrimination and both identification tapes. ABX discrimination tasks preceded the identification tasks. Within the tasks, half the subjects listened first to the sawtooth stimuli and then to the sine-wave stimuli, while the others listened in reverse order. They were not told about the labels pluck and bow until after the two discrimination tasks were completed. Otherwise, instructions were identical to Experiment I.

Results and Discussion

Sawtooth waves. As shown in the top panel of Figure 3, the identification function of the sawtooth stimuli is very quantal and the discrimination function quite peaked. Reversing the task order obviously had little, if any, effect; playing the stimuli over earphones did sharpen up both functions. The 0, 10, 20, and 30 msec stimuli were identified as plucked on 98 percent of all trials; the 50, 60, 70, and 80 msec stimuli were identified as bowed on 92 percent of all trials; and the 40 msec stimulus was ambiguous. The discriminability peak at the 30-50 msec comparison is significantly different from the 40-60 comparison (z = 2.74, $p < .005$), and while not significantly different from its other adjacent comparison, it is significantly different from 10-30 (z = 2.16, $p < .05$).

Sine waves. In the lower panel of Figure 3 are the identification and discrimination functions of the sine-wave stimuli. The identification function is considerably different from that of the sawtooth waves. While the 0, 10, 20, and 30 msec stimuli were identified as plucked on 94 percent of all trials, the 50, 60, 70, and 80 msec stimuli were identified as bowed on only 77 percent of all trials. The discrimination function, however, is quite peaked. As for the sawtooth waves, the 30-50 msec comparison for sine waves is significantly different from the comparisons 40-60 (z = 2.16, $p < .05$) and 10-30 (z = 2.45, $p < .01$).

A comparison of the musical stimuli. In terms of the discrimination results, categorical perception is clearly evident in both types of stimuli. Moreover, the category boundary as determined by the peak in the discrimination function is in the same place, about 40 msec. Although performance levels were slightly lower for the sine waves on the ABX task, there was no statistically significant difference between the nonspeech sounds.

One could hardly ask for a more striking identification function from the sawtooth stimuli. The sine-wave identification function, however, is sloppy at longer rise times. The identification curves for the sine waves do not differ significantly from those for the sawtooth stimuli at rise times of 0 through 50 msec. Beyond 50 msec, however, the results for the two waveforms diverge.

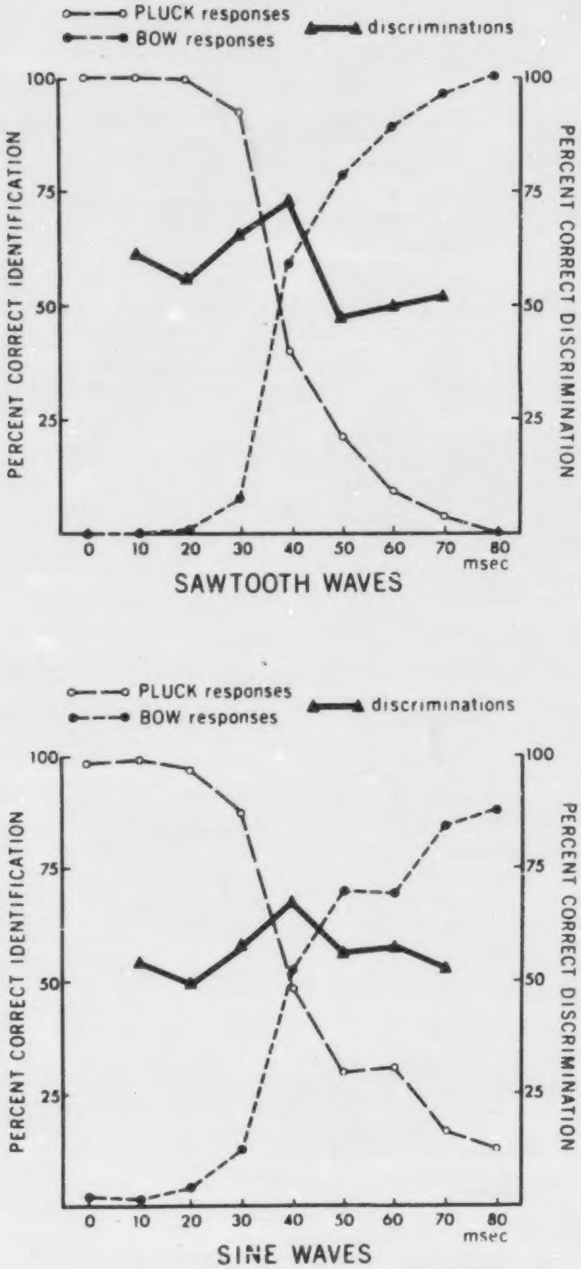


Figure 3: Identification and discrimination functions for sawtooth and sine waves.

Indeed, "bow" is not very impressive as a category for sine waves. Sine waves of long rise time sound more like a flute played legato style than like a stringed instrument. Thus, the category "bowed" may be inappropriate for sine waves. A better opposition might have been "staccato" for relatively short rise times versus "legato" for longer ones. The role of verbal labels in revealing categories may need careful attention.

Another possibility is that the continuum for sine waves of different rise times consists of only one category, "pluck" or perhaps "ping." If this seems odd, consider the following example. Imagine an acoustic continuum of stimuli at one end of which is a prototypic door slam. Imagine further that the stimuli in this continuum differ in rise time, just as in the present experiments. Since the acoustic invariant in door slams probably has something to do with rapid onsets, the more gradual the onset the less the waveform of particular items will resemble the door-slam prototype. One can imagine that at some point in the continuum the items no longer sound like door slams at all, and perhaps sound like nothing in this world. The question arises as to what lies beyond such a boundary: a noncategory, an "unnatural" category, or something else? An answer to this question, if indeed it is a real one, demands a fuller understanding than is available of the psychological nature and genesis of perceptual categories.

Categories and Boundaries in Audition

Categorical perception presents sharp contrasts to our more usual "continuous" perception. Most simple, psychophysical continua are perceived "continuously": one can discriminate many more items than he can identify (Pollack, 1952; Miller, 1956). In categorical perception, however, one may be able to discriminate only as well as he can identify the different items of the continuum. This unusual situation has been taken as a hallmark of speech perception (Liberman et al., 1967) and had been thought to occur in audition only for speech stimuli. We have found it to occur for nonspeech sounds, obviously demonstrating that it is not unique to speech. Adjustments need to be made in the general account of perceptual categories in audition. Nevertheless this finding should not be taken as disproof of categorical perception as a hallmark of speech. Instead, speech may be the domain of the most prominent exemplars of perceptual categories, even though such categories occur in both speech and nonspeech. In this discussion we shall try to develop this perspective.

The ABX task. First it is necessary to consider the nature and foibles of the ABX discrimination task and determine why it reveals categories and boundaries. The ABX task is peculiar in that it places the listener in a situation of echoic memory overload. Stimuli A and B must be remembered so that subsequent comparisons can be made with Stimulus X. Because echoic memory is simply insufficient to store all three stimuli, some form of coding must take place. Since Stimulus B arrives shortly after Stimulus A, A must be coded into some memorable form before B clobbers its echo; and B, in turn, must be coded before X clobbers it. By the time Stimulus X is in the system, Stimulus A may have been stripped of its acoustic husk, and Stimulus B may also have been coded into a memorable form. However, unlike A, B may also retain a wraith of an echo. Crucial within-category information may have been lost during coding of A and B simply because such differences were not in short-term memory. Therefore, within-category ABX judgments are often reduced to guessing propositions. The quick loss of the within-category information is the crux of categorical perception. Obviously, the ABX task is not the best possible discrimination task because of this memory and coding feature (Pisoni, 1971, 1973), but it may be most sensitive to perceptual categories.

Coding processes do not always involve this quick loss of acoustic information. In fact, our experiments are the first to demonstrate this for common nonlinguistic sounds. (For categorical perception of nonlinguistic sounds that simulate certain temporal aspects of speech, see Miller, Pastore, Wier, Kelly, and Dooling, 1974.) For stop consonants the acoustic nature of the speech sound is difficult to perceive: "the sound escapes us and we perceive the event, almost immediately, as phonetic" (Studdert-Kennedy, in press). It would seem that plucked and bowed sawtooth notes must share this nearly instantaneous coding in view of the results they yield in the ABX task.

However, a problem immediately arises in comparing these musical sounds to speech: plucked-note and bowed-note encoding cannot be phonetic. This fact, coupled with the result of the first experiment which demonstrated that rise time can cue perceptual categories in both speech and music, suggests that certain aspects of phonetic coding may be intimately related to the coding of naturally occurring nonlinguistic events.

Of speech codes and music codes. Liberman, Mattingly, and Turvey (1972) have noted that the speech code is an efficient code. Transforming an echo of a speech sound into a phonetic representation of that sound is roughly equivalent to transforming a 40,000 bit-per-second signal into a 40 bit-per-second signal. This enormous savings is at the cost of losing "unneeded" auditory information such as within-category differences. By the categorization process massive amounts of information in the auditory signal are transformed into "a unitary neural event" (p. 320). Such a unitary neural representation would obviously be easier to store and use for subsequent analyses and comparisons than would a degraded echo.

Our problem is then to explain why it is advantageous to code (categorize) plucks and bows. It seems exactly backwards to suppose that phonetic processes were "tricked" into analyzing our musical sounds as speech. An evolutionary view (Lieberman, 1973) is more reasonable. It assumes that speech perception developed around existing properties of the auditory system, one of which may be the ability to detect different categories along the dimension of rise time.

Are the two categories that lie astride the 40-msec rise time boundary innate to humans, and not learned? Two lines of evidence suggest a positive answer. First, training tends to increase discriminative capacity, but never to decrease it. Thus, the troughs at either side of the ABX discrimination peak cannot be explained by some pneumatic trade-off between the acquisition of the boundary and the loss of within-category discriminability (Liberman, Harris, Eimas, Lisker, and Bastian, 1961; Lane, 1965; Studdert-Kennedy, Liberman, Harris, and Cooper, 1970; Pisoni, 1971). Second, categorical perceptions of speech stimuli occur in one-month-old infants, and their data are functionally identical to those of adults (Eimas, Siqueland, Jusczyk, and Vigorito, 1971; Cutting and Eimas, in press). No one-month-old infant could have been tutored to perceive speech events in this manner. A crucial question arises: would infants perceive plucked and bowed notes categorically? We think they would, but it remains to be demonstrated. Because speech categories are not learned, we feel that pluck-and-bow perception may not be based on learning either.

How many categories in speech and in nonspeech? At this point we are drawn to compare the potential number of categories in speech and nonspeech. For a given language the count is easy: there are as many categories as there are phonemes to label them. In nonspeech the count is not so easy. Our ignorance keeps us from knowing where to look. Thus, perhaps the best yardstick is not to compare the number of categories, but to compare the number of different dimensions that can be manipulated to yield perceptions in different natural categories. For speech the list is impressive. Let us reconsider just the consonants. Differences in voicing for initial stop consonants (for example, [ba] versus [pa]) may be triggered by voice-onset time (Abramson and Lisker, 1965), by cutbacks in the first formant (Liberman, Delattre, and Cooper, 1958), and by inflections in the fundamental frequency (Haggard, Ambler, and Callow, 1970); in medial position by duration of closure (Lisker, 1957a); and in final position by vowel duration (Raphael, 1971). Differences in place of articulation for stops ([ba] versus [da] versus [ga]) may be triggered by the direction and extent of the second-formant transition (Liberman, 1957; Liberman et al., 1957) and of the third-formant transition (Harris, Hoffman, Liberman, Delattre, and Cooper, 1958), or by bursts (Liberman, Delattre, and Cooper, 1952). The third-formant transition is also a cue for liquids in initial position (O'Connor, Gerstman, Liberman, Delattre, and Cooper, 1957) and in medial position (Lisker, 1957b). Differences in fricatives may be triggered by the frequency of noise (Harris, 1958), duration of noise (Gerstman, 1957), or by transitions (Harris, 1958). Silent intervals within a syllable may cue differences between stops and semivowels (Bastian, Delattre, and Liberman, 1959), or between the presence and absence of a stop consonant (Bastian, Eimas, and Liberman, 1961). Other important cues are tempo (Liberman, Delattre, Gerstman, and Cooper, 1956), presence of nasal resonances (Liberman, Delattre, Cooper, and Gerstman, 1954), and of course rise time as shown by the present study and by Gerstman (1957). Not all manipulations of these parameters yield categorical perception, but many do, and this litany is by no means complete.

For common nonspeech sounds the list is not as impressive: so far, there is only rise time, as shown in the present paper. Are there other dimensions that cue differences between nonlinguistic categories? Perhaps, but the total number is not likely to approach that for speech. Some candidates which suggest themselves are frequency ratios for musical intervals and "steady-state" spectra for instrumental timbre.

Thus, in summary, we suggest that the presence of categories and boundaries should remain a hallmark of speech. Of course nonlinearities of perception occur in nonspeech as well, but that is as it should be. The fabric of speech perception and the mechanisms behind it could not have been woven wholly out of new cloth. Remnants of underlying auditory, nonlinguistic processes should, and do show through. The categorical perception of nonspeech stimuli varying in rise time is apparently one of these threads.

REFERENCES

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In Proceedings of the Fifth International Congress of Acoustics, Liege.

Bastian, J., P. C. Delattre, and A. M. Liberman. (1959) Silent interval as a cue for the distinction between stops and semivowels in medial position. J. Acoust. Soc. Amer. 31, 1568(A).

Bastian, J., P. D. Eimas, and A. M. Liberman. (1961) Identification and discrimination of a phonemic contrast induced by silent interval. J. Acoust. Soc. Amer. *33*, 842(A).

Cooper, F. S. and I. G. Mattingly. (1969) A computer-controlled PCM system for the investigation of dichotic speech perception. J. Acoust. Soc. Amer. *46*, 115(A).

Cutting, J. E. (in press) Different speech-processing mechanisms can be reflected in the results of discrimination and dichotic listening tasks. Brain and Language. [Also in Haskins Laboratories Status Report on Speech Research *SR-37/38* (this issue).]

Cutting, J. E. and P. D. Eimas. (in press) Phonetic feature analyzers and the processing of speech in infants. In the proceedings of the conference, "The Role of Speech in Language," ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press). [Also in Haskins Laboratories Status Report on Speech Research *SR-37/38* (this issue).]

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. Vigorito. (1971) Speech perception in infants. Science *171*, 303-306.

Fraisse, P. (1963) *The Psychology of Time*. (New York: Harper & Row).

Gerstman, L. J. (1957) Perceptual dimensions for the friction portion of certain speech sounds. Unpublished Ph.D. dissertation, New York University (Psychology).

Haggard, M. P., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. J. Acoust. Soc. Amer. *47*, 613-617.

Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. Lang. Speech *1*, 1-7.

Harris, K. S., H. S. Hoffman, A. M. Liberman, P. C. Delattre, and F. S. Cooper. (1958) Effect of third-formant transitions on the perception of voiced stop consonants. J. Acoust. Soc. Amer. *30*, 122-126.

Lane, H. (1965) Motor theory of speech perception: A critical review. Psychol. Rev. *72*, 275-309.

Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Amer. *29*, 117-123.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. *74*, 431-461.

Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1952) The role of selected stimulus variables in the perception of the unvoiced stop-consonants. Amer. J. Psychol. *65*, 497-516.

Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. Lang. Speech *1*, 153-167.

Liberman, A. M., P. C. Delattre, F. S. Cooper, and L. J. Gerstman. (1954) The role of consonant-vowel transitions in the perception of stop and nasal consonants. Psychol. Monogr. *68*, (8, Whole No. 379).

Liberman, A. M., P. C. Delattre, L. J. Gerstman, and F. S. Cooper. (1956) Tempo of frequency change as a cue for distinguishing classes of speech sounds. J. Exp. Psychol. *52*, 127-137.

Liberman, A. M., K. S. Harris, P. D. Eimas, L. Lisker, and J. Bastian. (1961) An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. Lang. Speech *4*, 175-195.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. *54*, 358-368.

Liberman, A. M., K. S. Harris, J. A. Kinney, and H. Lane. (1961) The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. *61*, 379-388.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington, D. C.: V. H. Winston).

Lieberman, P. (1973) On the evolution of language: A unified view. Cognition 2, 59-94.

Lisker, L. (1957a) Closure duration and the voiced-voiceless distinction in English. Language 33, 42-49.

Lisker, L. (1957b) Minimal cues for separating /w,r,l,y/ in intervocalic position. Word 13, 256-267.

Lisker, L. and A. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 1, 131-157.

Miller, G. A. (1956) The magical number seven, plus or minus two, or some limits on our capacity for processing information. Psychol. Rev. 63, 81-96.

Miller, J. D., R. E. Pastore, C. C. Wier, W. M. Kelly, and R. M. Dooling. (1974) Discrimination and labeling of noise-buzz sequences with varying noise-lead times. J. Acoust. Soc. Amer. 55, 390(A).

O'Connor, J. D., L. J. Gerstman, A. M. Liberman, P. C. Delattre, and F. S. Cooper. (1957) Acoustic cues for the perception of initial /w,j,r,l/ in English. Word 13, 25-43.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. dissertation, University of Michigan (Psycholinguistics). (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.

Pollack, I. (1952) The information of elementary multidimensional auditory displays. J. Acoust. Soc. Amer. 26, 155-158.

Raphael, L. J. (1971) Vowel duration as a cue to the perceptual separation of cognate sounds in American English. Ph.D. dissertation, City University of New York (Speech). (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Sawusch, J. R. and D. B. Pisoni. (1974) Category boundaries for speech and nonspeech sounds. J. Acoust. Soc. Amer. 55, 436(A).

Studdert-Kennedy, M. (in press) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (Springfield, Ill.: C. C Thomas).

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Different Speech-Processing Mechanisms Can be Reflected in the Results of Discrimination and Dichotic Listening Tasks*

James E. Cutting[+]
Haskins Laboratories, New Haven, Conn.

The relative peakedness of diotic ABX discrimination functions for certain speech stimuli and the relative magnitude of the right-ear advantage that such speech stimuli yield in dichotic listening tasks have often been thought to be functionally parallel measures of speech processing. The results of the present study suggest that this is not always the case.

The results of two very different experimental paradigms have been taken as primary evidence for distinguishing the perception of speech from the perception of other auditory events. These results are the discriminability functions associated with categorical perception, and the right-ear advantage in dichotic listening (see Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). Other auditory events, which for the most part can be handily classified as nonspeech, typically yield neither of these results (for an overview, see Studdert-Kennedy and Shankweiler, 1970; and Mattingly, Liberman, Syrdal, and Halwes, 1971). The most parsimonious explanation for why speech is processed in a unique manner in both experimental paradigms is that a similar mechanism or set of mechanisms underlies both results. Consider each in more detail.

Categorical perception and ABX discrimination. Usually a person can perceive and discriminate many more stimuli along a physical continuum than he can identify (Pollack, 1952, 1953). Often equal increments of physical change yield equal increments of perceptual change. However, for certain kinds of stimuli, especially certain kinds of speech stimuli, this is particularly untrue. The most thoroughly studied continua in speech perception have been the dimensions of place of articulation as manifested by second- and third-formant transitions (Liberman, Harris, Hoffman, and Griffith, 1957; Pisoni, 1971), and voice-onset time (Abramson and Lisker, 1965, 1970). Both are dimensions relevant to the perception of different stop consonants in most languages. Certain adjacent stimuli along these acoustic continua are difficult, if not impossible, to discriminate. Typically, subjects are able to discriminate items only as well as they can label them differently, and this labeling process is very nearly categorical. Thus, categorical perception implies a finely tuned ability to discriminate items that are acoustically similar but phonetically different, coupled with an inability to discriminate items that are acoustically similar but phonetically identical.

---

*To appear in Brain and Language.

[+] Also Yale University, New Haven, Conn.

Liberman, Harris, Kinney, and Lane (1961) have shown that the discriminability peaks associated with categorical perception occur for certain speech sounds but not for acoustically similar nonspeech sounds. Mattingly et al. (1971) have shown that these peaks occur for entire speech patterns, but not for their phonetically relevant acoustic cues when excised and presented in isolation as "chirps." Furthermore, Eimas and his colleagues (Eimas, Siqueland, Jusczyk, and Vigorito, 1971; Eimas, 1973, in press; Cutting and Eimas, in press) have shown that young infants perceive speech stimuli in a categorical manner and chirp stimuli in a more continuous manner, a result which is functionally identical to that of adults.

For the most part, however, the discrimination results have been found with adult subjects using ABX stimulus triads. On such trials the subject judges which of the first two stimuli in the triad is identical to the third stimulus. Pisoni (1971, 1973a, 1973b) has rigorously examined the processes underlying ABX discriminations, drawing on previous work by Fujisaki and Kawashima (1968, 1970). A perceptual model which stems from these papers includes both auditory and phonetic short-term memory components in order to explain the relative peaks and troughs in the discrimination function. The peaks appear to result from the engagement of the phonetic component of memory. The troughs, on the other hand, appear to reflect engagement of auditory memory. Performance here typically remains substantially above chance. The difference in performance between within-phoneme-boundary and across-boundary pairs may be interpreted as the difference in relative efficiency of the two types of memory. Thus, the shape of the discrimination function stems from both auditory and phonetic processes.

Right-ear advantages. In a dichotic listening situation, where one stimulus is presented to the right ear and another to the left ear, the subject is often unable to report all items that were presented on a particular trial. Whether both items are digits (Kimura, 1961), meaningful words (Bartz, Satz, Fennel, and Lally, 1967), or nonsense syllables (Shankweiler and Studdert-Kennedy, 1967), he is generally able to report more accurately the information presented to the right ear than to the left. This highly replicable result is called the right-ear advantage and has been explained, in part, in terms of the general processing capabilities of the cerebral hemispheres. Clinical evidence indicates that language processing occurs primarily in the left hemisphere for right handers (Milner, 1967; Geschwind, 1970), and that the crossed pathways from ear to cortex are more prominent during transmission than the uncrossed pathways (Puletti and Celesia, 1970). Thus, in the dichotic situation, stimuli presented to the right ear have privileged access to the speech processor in the left hemisphere. The manifestation of this privilege, then, is the right-ear advantage.

Certain aspects of the sound pattern of a speech utterance appear to require relatively more left-hemisphere processing than others. For example, Shankweiler and Studdert-Kennedy (1967) found that stop consonants in consonant-vowel (CV) nonsense syllables yielded a large right-ear advantage, while steady-state vowels yielded a considerably smaller, nonsignificant right-ear advantage. Other classes of phonemes such as liquids (Day and Vigorito, 1973; Crystal and House, 1974; Cutting, in press) often yield results intermediate between stops and vowels. To account for these results, Cutting (in press) has suggested that, along with a generalized speech processor in the left hemisphere, there may be an auditory analyzer whose task it is to compute frequency transitions and other purely auditory apsects of the signal. Indeed, Darwin (1971) found a right-ear advantage for fricatives with transitions in CV syllables and no ear advantage for fricatives without transitions. The data of Cutting (in press) suggest that this auditory device is engaged equally for speech and nonspeech signals. Thus, nonspeech signals which have quite a lot of transient information may occasionally yield right-ear advantages (Halperin, Nachshon, and Carmon, 1973).

The two phenomena compared. Although ABX discrimination tasks and dichotic listening tasks are quite different and appear to overload the perceptual system in very different ways, there are some impressive parallels between them. On empirical grounds, stop consonants yield sharply defined discrimination peaks and large right-ear advantages, whereas vowels typically yield less extreme results. (Nonspeech sounds, of course, often yield relatively flat discrimination functions and left-ear advantages.) On theoretical grounds, recent accounts of both phenomena have included both auditory and phonetic components. The present experiment was designed to compare the two phenomena in as direct a manner as possible. A specific attempt was made to devise a situation in which two tasks, ABX discrimination and dichotic recognition, would yield divergent results.

Method

Stimuli. Four arrays of seven stimuli each were synthesized on the Haskins Laboratories' parallel resonance synthesizer. All stimuli consisted of two formants, were 300 msec in duration, and had the same fundamental frequency (112 Hz). Two arrays consisted of CV syllables which began with either [b] or [d]. One array was synthesized with the vowel [a] and the other with [æ]. The steady-state formant frequencies for [a] were 743 Hz for the first formant (F1) and 1232 Hz for F2, while the corresponding values for [æ] were 743 Hz and 1620 Hz. Formant transitions were 50 msec in duration. The F1 transition was identical for both vowels, increasing linearly in frequency from a value of 437 Hz. Seven different F2 transitions were synthesized for each vowel: their initial values were 616, 769, 921, 1075, 1232, 1386, and 1541 Hz for the [ba]-to-[da] array; and 769, 921, 1075, 1232, 1386, 1541, and 1695 Hz for the [bæ]-to-[dæ] array. All F2 transitions were linear.

Two other arrays of stimuli were synthesized. They were identical to the CV stimuli except that the F1 transition was inappropriate for any particular phoneme segment. Instead of increasing in frequency for 50 msec, it decreased from a value of 894 Hz to 743 Hz for both vowels. The F2 transitions were identical to the seven used in the [ba]-[da] and [bæ]-[dæ] arrays. Since these stimuli resembled the CV stimuli along many dimensions, but did not have transitions corresponding to specific consonants, they were designated C'V stimuli. Schematic spectrograms of a CV and a C'V stimulus are shown in Figure 1, along with a display of the various possible F2 transitions.

In addition, an eighth stimulus was added to each of the four arrays: [ga] and [gæ] with very rapidly decreasing F2 transitions, and two C'V stimuli which corresponded to them.

ABX discrimination tapes. Members of each AB comparison were selected by pairing each stimulus with the item two steps removed along the F2 continuum; that is, stimulus 1 was paired with stimulus 3, 2 with 4, 3 with 5, 4 with 6, and 5 with 7, yielding five possible pairs. For each pair there were four possible ABX comparisons: ABA, ABB, BAB, and BAA. Two different random sequences of 80 items were prepared: (5 ABX pairs) X (4 ABX comparisons per pair) X (4 arrays of

FIGURE 1

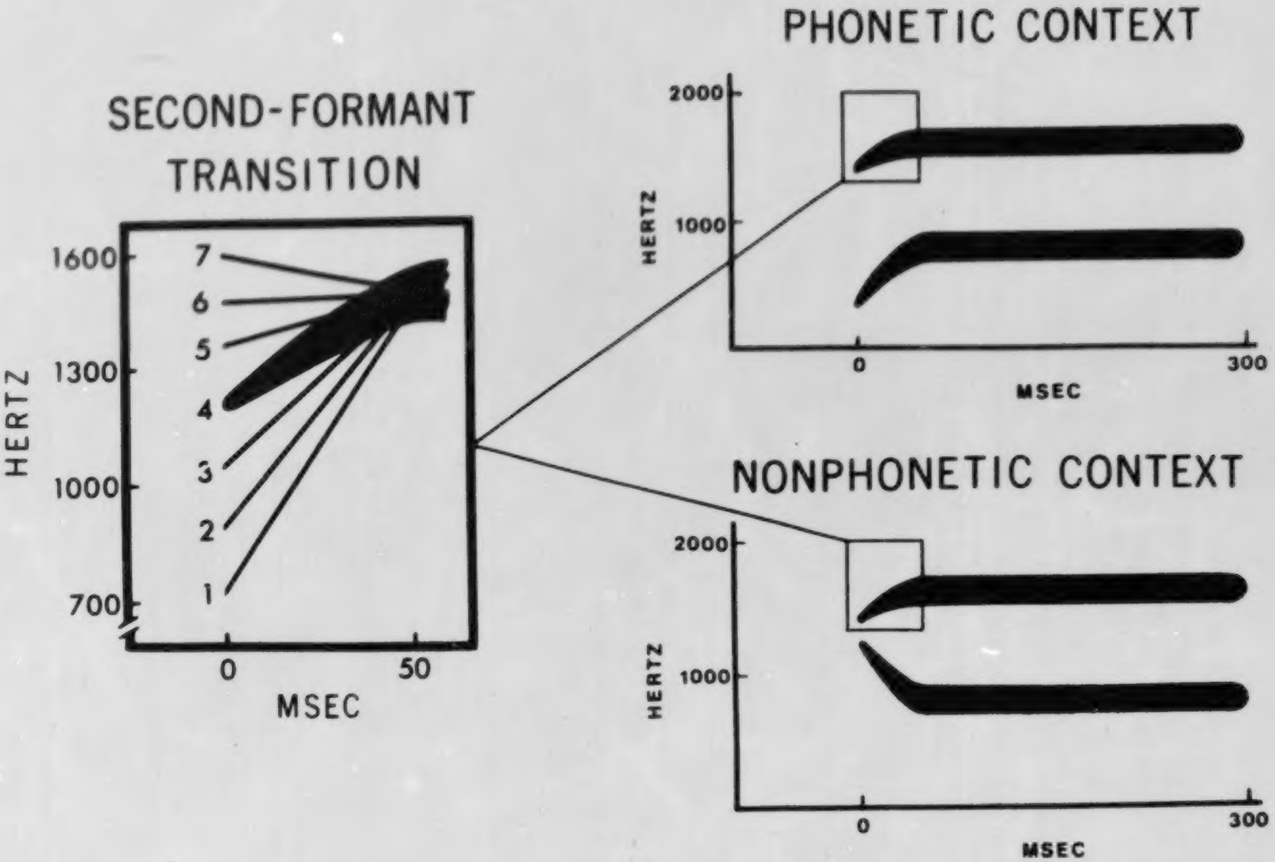

SECOND-FORMANT TRANSITION
HERTZ
1600
1300
1000
700
7
6
5
4
3
2
1
0
50
MSEC
PHONETIC CONTEXT
HERTZ
2000
1000
0
300
MSEC
NONPHONETIC CONTEXT
HERTZ
2000
1000
0
300
MSEC

Figure 1: Schematic spectrograms of stimuli with speech-relevant and speech-irrelevant first-formant transitions, and a display of the array of their second-formant transitions.

stimuli). The members of each ABX triad were separated by 1 sec, with 4 sec between triads.

Dichotic recognition tapes. Stimuli numbers 2 and 6 were selected from each of the four seven-item arrays used in the ABX task. Added to them was each of the eighth stimuli, yielding the CVs [ba, da, ga, bæ, dæ, gæ] and the corresponding C'Vs. Trials consisted of a dichotic pair, followed by 1 sec of silence, followed by a diotic probe. Dichotic pairs of CV stimuli were matched such that items shared neither the same consonant nor the same vowel: thus, [ba] was paired with [dæ] or [gæ], and [bæ] with [da] or [ga]. A similar rule was applied to the C'V pairs in that items shared neither the same vowel nor the F2 transition normally associated with a particular stop in that vowel context. On half the trials the diotic probe was one of the stimuli presented to the left ear for one quarter of the trials and to the right ear for the other quarter, and on half the trials the probe was a third stimulus not previously presented. Two different random sequences of 96 items were prepared: (6 pairings per stimulus class) X (2 classes of stimuli - CV and C'V) X (2 channel arrangements) X (2 possible legitimate probes) + 48 similar trials in which the probe was not a member of the dichotic pair.

Subjects and apparatus. Sixteen Yale University undergraduate students participated in two tasks, diotic ABX discrimination and dichotic recognition. All subjects were right-handed, native American English speakers with no history of hearing difficulty, and no previous experience at dichotic listening or at listening to synthetic speech. Audio tapes were played on an Ampex AG500 dual-track tape recorder, and signals were sent through a listening station to Grason-Stadler earphones (Model TDH39-300Z).

Procedure. Before the ABX task began, subjects were instructed to write down which stimulus, A or B, was identical to the third member of the triad. They listened to four practice trials to familiarize themselves with the task and the stimuli. Before the dichotic recognition task began, they were instructed to attend to one ear for a block of trials, and to write down Y for yes if they thought the probe had been presented to the attended ear, and N for no if it was not. Counterbalancing of the monitored ear was done within subjects, and counterbalancing of ear-to-channel assignments was done across subjects. They listened to six practice trials before monitoring a given ear. Half of the subjects participated in the discrimination task before the dichotic recognition task, while the others participated in reverse order.

Results

Discrimination. CV stimuli yielded results typical of those associated with categorical perception: the discriminability of 3-5 pairs averaged 80 percent correct, much superior to all other CV pairs. Stimulus 3 was typically heard as [ba] or [bæ] and Stimulus 5 as [da] or [dæ] according to preliminary identification results. Thus these stimuli unambiguously belonged to different phoneme categories. C'V stimuli did not yield any discrimination peak; all comparisons were within a few percentage points of 60 percent correct. In general subjects were better at discriminating CV stimuli than C'V stimuli ($F(1,15) = 27.9$, $p < .001$), but this superiority appeared to be largely a result of the interaction of the shapes of the two discrimination functions, as shown in Figure 2. The interaction of the array of two-step comparisons with the stimulus classes, CV and C'V, was significant ($F(4,15) = 4.5$, $p < .025$). Of 16 subjects, 13 discriminated

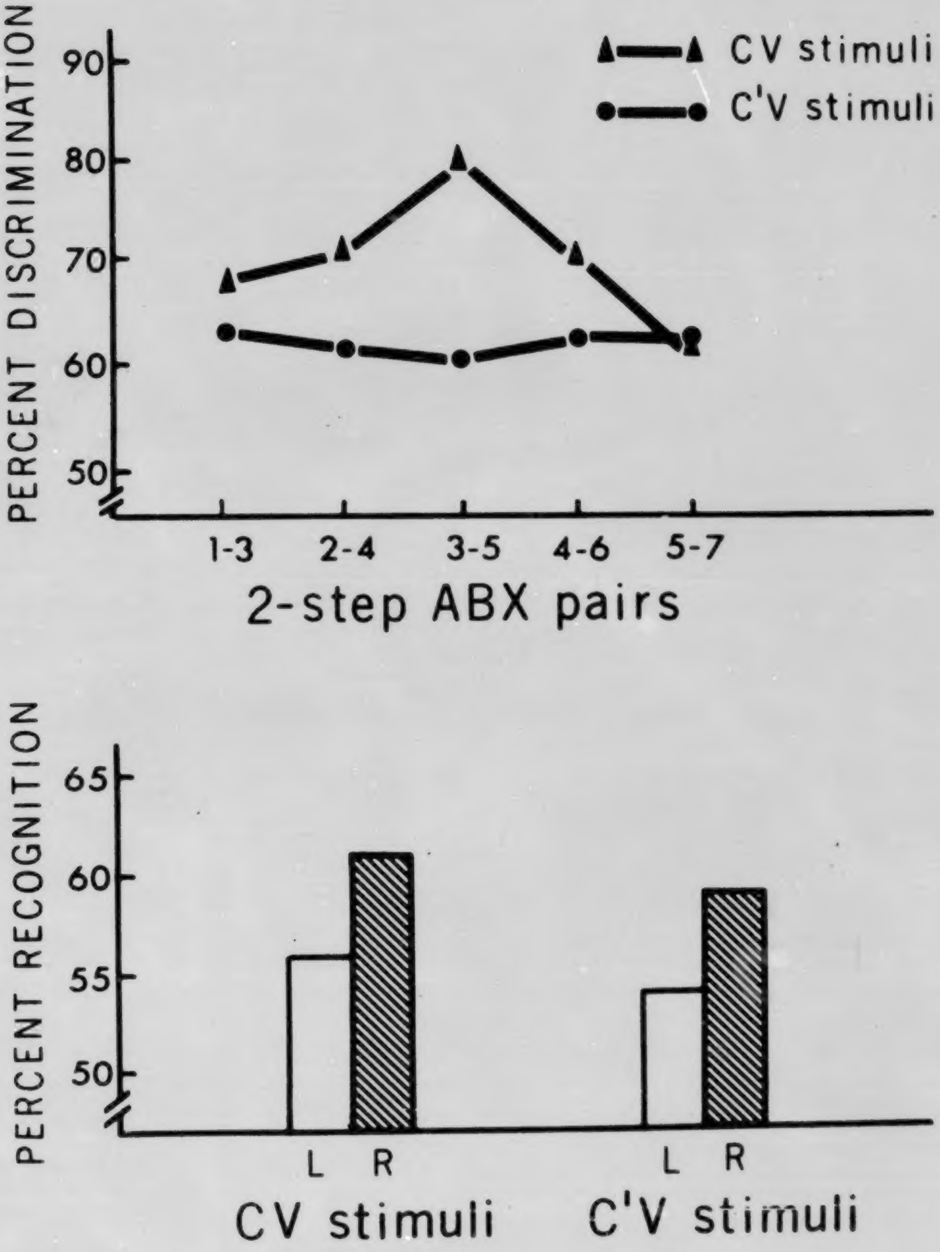


Figure 2: The results of the ABX discrimination task and the dichotic recognition task.

3-5 CV pairs better than 3-5 C'V pairs, and no subject performed better on the C'V comparison. There was no significant difference between the CV and C'V pairs for any other two-step comparison. Furthermore, the discrimination functions shown in Figure 2 were typical of all 16 subjects.

Dichotic recognition. CV and C'V stimuli yielded nearly identical right-ear advantages, as shown at the bottom of Figure 2. For the CV pairs, subjects were 61 percent correct at recognizing the attended stimulus in the right ear, while only 57 percent correct at recognizing that stimulus in the left ear. The corresponding scores for the C'V stimuli were 59 and 55 percent. Thus, both classes of stimuli yielded a 4 percent right-ear advantage ($F(1,15) = 5.35$, $p < .05$). There was no significant difference in the two ear advantages, nor in their levels of overall performance. Neither ear advantage, however, was significant by itself.

Discussion

The results of the present study suggest that the relative peakedness of discrimination functions normally associated with categorical perception and the relative magnitude of the right-ear advantage in dichotic listening are not functionally parallel results. Furthermore, in certain situations the mechanisms that underlie them cannot be exactly the same. Before considering the ramifications of this finding, however, it is necessary to consider the stimuli, paradigms, and results in greater depth.

C'V stimuli are essentially unidentifiable in any way other than a purely arbitrary fashion. This is partly because they could never have been produced by a human vocal tract. On the other hand, CV stimuli, like all speech items, have the peculiar feature of naming themselves--that is, providing their own nonarbitrary label--precisely because they could be produced by any normal human vocal tract. This difference between the two classes of stimuli determined the nature of the dichotic listening task. Most dichotic tasks which employ speech stimuli are tasks in which subjects identify, in an oral or written form, the stimuli they heard. Because the C'V stimuli could not be reproduced by the subjects, in either oral or written forms, it was necessary to devise either a recognition task or an identification task which used arbitrary labels for the C'V items. The recognition task was selected to avoid introducing the variable of differential labeling into the subjects' responses. Furthermore, the task was similar to that used by Kimura and Folb (1968) and Spellacy and Blumstein (1970). No special modification, of course, was necessary for the discrimination task.

The results of the dichotic recognition task showed that both CV and C'V stimuli yielded right-ear advantages. Although significant when considered in conjunction, neither ear advantage was significant by itself. Ideally, one would like both ear advantages to be significant, to demonstrate that the left hemisphere was superior in processing each kind of stimulus. Nevertheless, the fact that the ear advantages were nearly identical in magnitude and that taken together they did yield a significant ear advantage is a compelling result. That the ear advantages are small is not unusual. Small ear differences appear to be a hallmark of recognition tasks (see again, the results of Kimura and Folb, 1968; Spellacy and Blumstein, 1970).

Perceptual mechanisms. The mechanisms that underlie the results of the discrimination task are quite transparent. The peak in the CV function appears to be the result of the engagement of the phonetic component of short-term memory. The absence of a peak in the C'V functions at any position appears to be a result of the absence of any opportunity for phonetic memory coding to occur. Instead, the above-chance performance for all five C'V comparisons and for the endpoint CV comparisons appears to reflect the relative strength of memory for purely auditorily coded information.

The mechanisms that underlie the results of the dichotic recognition task are less transparent. There are two candidates: a generalized speech processor and an auditory analyzer. Cutting (in press) has shown that the left hemisphere system appears to excel its counterpart not only in processing speech, but also in processing some purely auditory aspects of the signal. The two mechanisms appear to be independent and additive. In the present study, however, it is not clear whether a phonetic mechanism, an auditory mechanism, or both account for the results. Consider these three possibilities in reverse order.

First, CV and C'V dichotic pairs may have yielded nearly identical results because they invoked both auditory and phonetic left-hemisphere mechanisms to a similar extent. After all, both classes of stimuli contained formant transitions which would need to be analyzed, and both had vowel segments upon which basic phonetic decisions could be made. This explanation is based on the assumption that C'V stimuli are speech stimuli--an assumption that is, I think, entirely valid. C'V stimuli sound like speech syllables with a very garbled beginning. The "garbledness" of the stimulus onsets result from inappropriate formant transitions. Nonetheless, backwards speech stimuli, such as those used by Kimura and Folb (1968), could be no less garbled and yet they too yield a right-ear advantage. The major argument against this explanation is that the ear advantages in the present study are quite small, perhaps too small to be convincing evidence of two components underlying them. The data of Cutting (in press) show that CV and C'V stimuli yield much larger right-ear advantages in a temporal-order judgment task.

Second, perhaps both classes of stimuli, CV and C'V, invoked only the phonetic decision-making mechanism in the left hemisphere. Since members of each dichotic pair differed in both transitions and vowels, subjects may have used only the vowel dimension of the stimuli to base their judgments on. Although this explanation appears to be tenable, it seems unlikely since Cutting (in press) found in both identification and temporal-order judgment tasks that dichotic speech stimuli that differed in transitions and vowels yielded larger right-ear advantages than pairs of different steady-state vowels. Such a result strongly implies that transitions are processed in this situation, and that they contribute to the ear advantage.

Third, and perhaps most likely, both classes of stimuli may have invoked only auditory processing mechanisms, and the results may reflect only the superiority of the left-hemisphere system for this type of auditory perception and memory. The task, after all, did not require phonetic identifications. Speech processing per se may not have been involved at all. Day and her coworkers (Day and Cutting, 1970; Day, Cutting, and Copeland, 1971; Wood, Goff, and Day, 1971; Day, in press) have shown that the speech processor can be disengaged in a variety of dichotic and diotic tasks, lending credence to this explanation. Indeed, perhaps the reason that the ear advantages in the present study were comparatively small in relation to those of other studies (see Studdert-Kennedy and Shankweiler, 1970; Cutting, in press) is that only a left-hemisphere auditory processor was engaged. Cutting (in press) has suggested that the increment in right-ear advantage due to auditory processing is less than that due to phonetic coding, or speech processing in the traditional sense.

Summary and Conclusion

Evidence from many discrimination studies and dichotic listening studies has suggested that there is a functional parallel between the relative sharpness of discrimination curves and the relative magnitude of the right-ear advantage. The present study demonstrates that this is not always the case. Speech discrimination functions appear to be a result of both auditory and phonetic processing, whereas a right-ear advantage may result from the engagement of one or both types of processing mechanisms.

REFERENCES

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In *Proceedings of the Fifth International Congress of Acoustics*, Liege.

Abramson, A. S. and L. Lisker. (1970) Discriminability along the voicing continuum: Cross-language tests. In *Proceedings of the Sixth International Congress of Phonetic Sciences*, Prague, 15-25.

Bartz, W. H., P. Satz, E. Fennel, and J. R. Lally. (1967) Meaningfulness and laterality in dichotic listening. J. Exp. Psychol. *73*, 204-210.

Crystal, T. H. and A. S. House. (1974) Effect of local signal level on differential performance in dichotic listening. J. Acoust. Soc. Amer. *55*, 434(A).

Cutting, J. E. (in press) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys.

Cutting, J. E. and P. D. Eimas. (in press) Phonetic feature analyzers and the processing of speech in infants. In the proceedings of the conference, "The Role of Speech in Language," ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press). [Also in Haskins Laboratories Status Report on Speech Research *SR-37/38* (this issue).]

Darwin, C. J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. *23*, 46-62.

Day, R. S. (in press) Engaging and disengaging the speech processor. In *Hemispheric Asymmetry of Function*, ed. by M. Kinsbourne. (London: Tavistock).

Day, R. S. and J. E. Cutting. (1970) Levels of processing in speech perception. Paper presented at the 10th Annual Meeting of the Psychonomic Society, San Antonio, Tex., November.

Day, R. S., J. E. Cutting, and P. M. Copeland. (1971) Perception of linguistic and nonlinguistic dimensions of dichotic stimuli. Paper presented at the 11th Annual Meeting of the Psychonomic Society, St. Louis, Mo., November.

Day, R. S. and J. M. Vigorito. (1973) A parallel between encodedness and the ear advantage: Evidence from a temporal-order judgment task. J. Acoust. Soc. Amer. *53*, 358(A).

Eimas, P. D. (1973) Linguistic processing of speech by young infants. Paper presented at the conference "Language Intervention with the Mentally Retarded," Wisconsin Dells, Wis., June.

Eimas, P. D. (in press) Speech perception in early infancy. In *Infant Perception*, ed. by L. B. Cohen and P. Salapatek. (New York: Academic Press).

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. M. Vigorito. (1971) Speech perception in infants. Science *171*, 303-306.

Fujisaki, H. and T. Kawashima. (1968) The influence of various factors on the identification and discrimination of synthetic speech sounds. In *Reports of the Sixth International Congress on Acoustics*, Tokyo, August.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute (Faculty of Engineering, University of Tokyo) 29, 207-214.

Geschwind, N. (1970) The organization of language and the brain. Science 170, 940-944.

Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift in ear superiority in dichotic listening to temporal pattern nonverbal stimuli. J. Acoust. Soc. Amer. 53, 46-50.

Kimura, D. (1961) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. and S. Folb. (1968) Neural processing of backwards-speech sounds. Science 161, 395-396.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Liberman, A. M., K. S. Harris, J. A. Kinney, and H. L. Lane. (1961) The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. 61, 379-388.

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Milner, B. (1967) Brain mechanisms suggested by studies of temporal lobes. In Brain Mechanisms Underlying Speech and Language, ed. by C. H. Millikan and F. L. Darley. (New York: Grune and Stratton) 122-145.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. thesis, University of Michigan (Psycholinguistics). (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Pisoni, D. B. (1973a) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.

Pisoni, D. B. (1973b) The role of auditory short-term memory in vowel perception. Haskins Laboratories Status Report on Speech Research SR-34, 89-118.

Pollack, I. (1952) The information in auditory display. J. Acoust. Soc. Amer. 24, 745-749.

Pollack, I. (1953) The information in elementary auditory display. II. J. Acoust. Soc. Amer. 25, 765-769.

Pulletti, F. and G. G. Celesia. (1970) Functional properties of the primary cortical auditory area in man. J. Neurosurg. 32, 244-247.

Shankweiler, D. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Spellacy, F. and S. Blumstein. (1970) The influence of language set on ear preference in phoneme recognition. Cortex 6, 430-439.

Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Wood, C. C., W. R. Goff, and R. S. Day. (1971) Auditory evoked potentials during speech perception. Science 173, 1248-1251.

The Intelligibility of Synthetic Monosyllabic Words in Short, Syntactically Normal Sentences.

P. W. Nye and J. H. Gaitenby
Haskins Laboratories, New Haven, Conn.

INTRODUCTION

As a phase in Haskins Laboratories' development of high quality synthetic speech for use by blind "readers," an evaluation study is under way to assess the intelligibility of the current synthetic speech. The present test employed a series of meaningless, but syntactically normal sentences constructed from frequently used English monosyllabic words. This test will be referred to as the SNST (Syntactically Normal Sentence Test). In an earlier phase of the study, the well-known Modified Rhyme Test (MRT) was used to identify those consonants (in isolated words) that produced significant substitution errors. However, a number of shortcomings in the MRT format become evident as the data analysis progressed (Nye and Gaitenby, 1973). First, although the MRT was successful in identifying poorly synthesized phones, the particular consonantal substitutions made were of doubtful relevance for the purpose of correcting the synthesis rules. The reason for this limitation lay in the closed-response design of the MRT which forced the subjects to choose one of six rhyming words as the match to the stimulus heard. Frequently the alternative consonants provided in each rhyming choice set were phonetically too distant from the intended stimulus consonant to be a reasonable substitution. The MRT also suffered from the fault that not all the English consonants were represented in their customary initial and final positions, and that some consonants did not appear at all. Moreover, vowel environments were also unbalanced and incomplete. Nevertheless, the MRT did provide some useful data, making possible a preliminary comparison of subjects' performances with synthetic and natural speech. The overall MRT intelligibility scores were found to be 92.4 percent for synthetic speech, and 97.3 percent for the parallel test in natural speech.

The Objective and General Design Features of the SNST

The two-fold purpose of the SNST was to appraise the quality of the present synthetic speech and to pinpoint the acoustic features and phonetic classes of

---

Acknowledgment: Research support to Haskins Laboratories for this work was provided by the Prosthetics and Sensory Aids Service of the Veterans Administration, with the assistance of a University of Connecticut Research Foundation Grant to Dr. J. David Hankins of the University. We thank Dr. Hankins for his assistance and also wish to express appreciation to Lea Donald and Margaret Allen for their part in administering the tests and tallying the data.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

those phonemes producing high substitution error--in order to specify necessary synthesis rule revision. The main design constraints and considerations were that:

(a) real words were to be used as stimuli in order to obtain written responses from naive listeners;

(b) open (rather than closed) responses would be required; and

(c) the test stimuli would be presented in connected speech rather than in isolation.

Among the other considerations which influenced the design of the SNST were that:

(d) the intelligibility of both consonants and vowels should be tested in proportions representative of English;

(e) the words in the SNST should be frequently used monosyllables (thus presenting a more difficult identification task than longer words would pose);

(f) the words should be employed in a sentence-like environment, with normal syntax, but semantically meaningless in order to reduce word identification cues; and

(g) a parallel test using the same material in natural speech should be administered to all listeners to establish the highest standard against which the data on synthetic speech could be compared.

The new test thus represented a considerable advance in difficulty beyond the MRT. Its open-response design brought into play the interaction of prosodic and allophonic features with intelligibility and memory factors.

METHOD

Construction of the SNST

Monosyllabic words were drawn from the 2,000 most frequently used English words in the Thorndike and Lorge (1968) word count. The word list included 63 adjectives, 63 verbs in the past tense, and 126 nouns. Sentences were constructed by selecting words on a pseudo-random basis from the appropriate grammatical classes in the order: The (adjective) (noun) (verb, past tense) the (noun). The two nouns received high stress (raised pitch and extended duration), the verbs and the adjectives received mid stress (low pitch, extended duration), and the word "the" received low stress (low pitch, short duration). A full list of the sentences used is given in the four series of Table 1. Each sentence was generated by Haskins Laboratories' parallel formant resonance synthesizer--using the synthesis-by-rule program, which provided both the phones and the generally descending fundamental frequency characteristics of a statement.

Two-hundred sentences, divided into four series of 50 sentences, were synthesized at a speaking rate of 130 words per minute. A human speaker (AA) then listened to each synthesized sentence in turn and carefully repeated it at

TABLE 1: **Series 1**

1. The wrong shot led the farm.
2. The black top ran the spring.
3. The great car met the milk.
4. The old corn cost the blood.
5. The short arm sent the cow.
6. The low walk read the hat.
7. The rich paint said the land.
8. The big bank felt the bag.
9. The sick seat grew the chain.
10. The salt dog caused the shoe.
11. The last fire tried the nose.
12. The young voice saw the rose.
13. The gold rain led the wing.
14. The chance sun laid the year.
15. The white bow had the bed.
16. The near stone thought the ear.
17. The end home held the press.
18. The deep head cut the cent.
19. The next wind sold the room.
20. The full leg shut the shore.
21. The safe meat caught the shade.
22. The fine lip tired the earth.
23. The plain can lost the men.
24. The dead hand armed the bird.
25. The fast point laid the word.
26. The mean wave made the game.
27. The clean book reached the ship.
28. The red shop said the yard.
29. The late girl aged the boat.
30. The large group passed the judge.
31. The past knee got the shout.
32. The least boy caught the dance.
33. The green week did the page.
34. The live cold stood the plant.
35. The third air heard the field.
36. The far man tried the wood.
37. The high sea burned the box.
38 The blue bill broke the branch.
39. The game feet asked the egg.
40. The ill horse brought the hill.
41. The strong rock built the ball.
42. The dear neck ran the wife.
43. The dry door paid the race.
44. The child share spread the school.
45. The brown post bit the ring.
46. The clear back hurt the fish.
47. The round work came the well.
48. The good tree set the hair.
49. The bright guide knew the glass.
50. The hot nest gave the street.

TABLE 1: Series 2

51. The new wife left the heart.
52. The mean shade broke the week.
53. The hard blow built the truth.
54. The next game paid the fire.
55. The first car stood the ice.
56. The hot box paid the tree.
57. The live farm got the book.
58. The white peace spoke the share.
59. The black shout caught the group.
60. The end field sent the point.
61. The sick word had the door.
62. The last dance armed the leg.
63. The fast earth lost the prince.
64. The gray boat bit the sun.
65. The strong ring shot the nest.
66. The rich branch heard the post.
67. The gold glass tried the meat.
68. The dark cow laid the sea.
69. The deep shoe burned the face.
70. The north drive hurt the dog.
71. The chance wood led the stone.
72. The young shore caused the bill.
73. The least lake sat the boy.
74. The big hair reached the head.
75. The short page let the knee.
76. The bad bed said the horse.
77. The bright cent caught the king.
78. The fine bag ran the car.
79. The old fish called the feet.
80. The late milk made the cold.
81. The clear well asked the air.
82. The dear hill tried the work.
83. The full plant cut the voice.
84. The game boy thought the back.
85. The east floor brought the home.
86. The brown chair paid the girl.
87. The plain drink cost the wind.
88. The dark road net the hold.
89. The new truth sat the blow.
90. The gray prince called the hall.
91. The march face spoke the peace.
92. The hard heart let the bay.
93. The north king paid the drive.
94. The first oil put the drink.
95. The light eye hurt the lake.
96. The bad ice beat the floor.
97. The best house left the floor.
98. The east show found the cloud.
99. The cool lord paid the grass.
100. The coarse friend shot the chair.

TABLE 1: Series 3

101. The march hall aged the neck.
102. The great cloud read the road.
103. The past egg passed the shot.
104. The round blood grew the wind.
105. The cool rose spread the eye.
106. The light ball held the bow.
107. The salt wing tired the oil.
108. The low net set the show.
109. The large year ran the bank.
110. The red school hurt the house.
111. The near bird did the can.
112. The third press met the arm.
113. The blue race shut the rock.
114. The ill land put the friend.
115. The green chain knew the man.
116. The coarse judge saw the walk.
117. The safe hat felt the lord.
118. The child yard laid the hand.
119. The dry gate found the wave.
120. The best nose gave the corn.
121. The good grass held the paint.
122. The high street said the top.
123. The wrong room sold the rain.
124. The far ship beat the guide.
125. The right spring led the seat.
126. The wrong head thought the farm.
127. The black corn sent the word.
128. The strong prince came the grass.
129. The short boy paid the school.
130. The dark share hurt the earth.
131. The north friend gave the drink.
132. The dead book grew the plant.
133. The clean show left the men.
134. The safe knee paid the rose.
135. The far voice called the ring.
136. The march oil asked the peace.
137. The last tree did the egg.
138. The next eye shot the ball.
139. The salt bill broke the dance.
140. The fine truth tired the ear.
141. The white sun got the boat.
142. The coarse paint shut the bird.
143. The red back said the hold.
144. The least can sold the chair.
145. The end rock lost the shoe.
146. The sick neck led the hat.
147. The green ice passed the hill.
148. The big bow spread the lake.
149. The late point sat the branch.
150. The great leg armed the milk.

TABLE 1: Series 4

151. The brown bank tired the floor.
152. The deep shop sold the dance.
153. The gold truth cost the ball.
154. The big work burned the bird.
155. The last arm hurt the shade.
156. The low walk lost the nose.
157. The blue eye broke the plant.
158. The fast face grew the shoe.
159. The large home caused the ear.
160. The rich wave beat the net.
161. The light post held the field.
162. The dark bill left the branch.
163. The best man felt the gate.
164. The dear work met the ship.
165. The ill seat read the cent.
166. The live home caught the spring.
167. The round shot laid the shout.
168. The hot door heard the bed.
169. The brown lord tried the cow.
170. The mean arm spoke the land.
171. The large hand burned the game.
172. The blue nest aged the bay.
173. The past horse made the shade.
174. The hard girl caused the blood.
175. The game road found the page.
176. The third stone said the net.
177. The young air had the rose.
178. The dry wind laid the floor.
179. The bright dog saw the glass.
180. The bad house hurt the hair.
181. The gray car knew the wood.
182. The fast lip ran the field.
183. The first wave built the yard.
184. The gold walk let the box.
185. The clear shop cost the ball.
186. The low king bit the wing.
187. The cool sea led the bag.
188. The old guide beat the well.
189. The child top put the shore.
190. The rich group stood the press.
191. The high five set the chain.
192. The east face paid the judge.
193. The plain post tried the cloud.
194. The chance bank caught the blow.
195. The full week reached the race.
196. The deep heart cut the year.
197. The good cold held the wife.
198. The near rain sang the drive.
199. The new feet brought the street.
200. The light meat ran the fish.

essentially the same speaking rate with vocal inflection as close as possible to the synthetic sentence. When the natural speech recordings had been completed, the sentences of series 2 and 4 were rearranged in the reverse order to that used in the synthetic speech recordings, to provide a counterbalance against order bias.

Experimental Subjects

Four groups of eight listeners were assembled from student volunteers at the University of Connecticut. All 32 subjects were paid for their participation. Two of the groups consisted of listeners who had taken part in the Modified Rhyme Test, had achieved above-average scores, and had already made progress toward accommodating themselves to synthetic speech. The remaining 16 subjects, forming two more groups, were totally inexperienced either at participating in listening tests or at listening to synthetic speech. These "Old" and "New" subjects were selected to permit a study of the distribution of errors made by inexperienced subjects in comparison with errors by subjects who were somewhat used to synthetic speech.

Before the actual tests, each subject was given a hearing test and was then presented with ten synthetic speech sentences which could be listened to repeatedly in order to gain familiarity with the presentation procedure.

Mode of Presentation

Each subject received response sheets on which dotted lines (The .... the .....) were provided to indicate where the words heard should be entered. A 10-sec interval elapsed between the presented sentences to allow the listeners time to write their responses.

As in the earlier series of tests, the sentences were heard in a sound-damped booth through pairs of Grason-Stadler binaural earphones (type TDH39-300Z with earmuffs) at an approximate speaking level of 80 db SPL measured on output of the vowel /æ/. Table 2 shows the order of presentation followed during the experiment.

TABLE 2: Order of presentation for the Syntactically Normal Sentence Test.

GROUP	RUN	RUN	RUN	RUN	RUN	RUN	RUN	RUN
A*	S1	N2	S3	N4	S2	N1	S4	N3
B*	N3	S4	N1	S2	N4	S3	N2	S1
C	S2	N1	S4	N3	S1	N2	S3	N4
D	N4	S3	N2	S1	N3	S4	N1	S2

KEY: * indicates groups containing experienced subjects obtained from the previous experiment. The two-character identification code indicates whether the sentences are presented in synthetic (S) or natural (N) speech and the number indicates the particular series (1-4).

Each series of 50 sentences was heard once in synthetic speech (at one session) and once in natural speech (at another session). The order of presentation was varied among listener groups to avoid order biases.

Data Analysis

As expected, many more errors were made in reporting whole sentences than had been found with the single-word stimuli in the MRT. The SNST was far more difficult because it required the subject to recall and write down all four key words of each sentence, and some of the errors that occurred were more likely to have been caused by problems of memory than by phonetic ambiguities. Hence, words were sometimes left blank or reported in the wrong sequence. Also, occasional words were reported with more pnonemes than had been intended, or with fewer phonemes.

Acknowledging these facts, the analysis procedure involved a breakdown of the phonetic errors into three categories, and errors at the word level into two categories.

Phonetic Errors

(a) Substitutions - the substitution of a vowel or consonant by another, e.g., fat for sat, sat for sad, said for sad.

(b) Insertions - the insertion of one or two vowels or consonants in an otherwise correctly reported word, e.g., paved for paid

(c) Deletions - the omission of a vowel or consonant in an otherwise correctly reported word, e.g., paid for paved.

Word Errors

(a) Omissions - words left unreported.

(b) Transpositions - words reported correctly but in the wrong position within the sentence.

The computer coding scheme that was developed allowed each error and its location to be recorded as response sheets were checked. These data were later sorted by a computer program to yield error lists for each category. In the case of phonetic errors, the sorting routine also provided a list of the phones that preceded and followed each error.

RESULTS

Natural Speech Results

The absolute yardstick against which the performance with synthetic speech can be measured is the data obtained from the parallel natural speech tests. Such a comparison sets the highest possible standard for synthetic speech--a standard that may not be necessary to reach for the practical purposes of a high-speed reading machine service, at least initially. Nevertheless, it is appropriate to begin with a review of the natural-speech results, illustrated in Figure 1.

Both groups of subjects (Old and New) performed similarly on the natural-speech tests and their data have therefore been pooled. A total of the errors of all kinds (ranging from words completely omitted to minor phonetic errors--all counted as whole word errors) yielded an average error rate of about 5 percent on a base of 800 words presented. This was higher than the overall error rate of 3 percent found in the MRT but, as has been noted already, the present test was more difficult and a higher error rate was to be expected.

No phonemes produced more than 2 percent substitution errors in syllable-final position. In initial position, however, two phonemes produced confusions exceeding 2 percent: /θ/ (3.9 percent) and /š/ (3.3 percent). The phoneme /θ/ was heard chiefly as /f/, and /š/ was heard as /č/. Thus, both of these voiceless sounds produced Place confusions, and /š/ produced a Manner error as well. (In final position /š/ produced no confusions at all, and the /θ/ confusions fell below 1 percent.) The vowel /ɝ/ produced a 2.6 percent error.

Synthetic Speech Errors

Errors in the synthetic speech test occurred at a substantially higher rate than was observed in the natural speech data. The overall average error rate (again including errors of all types) was 22 percent compared with the 8 percent rate for synthetic speech errors obtained from the MRT. Some of the SNST errors, Omissions and Transpositions in particular, were more likely to have been caused by memory lapses than by phonetic ambiguities. The greatest discrepancy between the two groups of subjects was in the Omissions category, in which the Old subjects failed to respond to 1.6 percent of the synthetic words presented, while the New subjects omitted responses to 8.1 percent. Word Transpositions, on the other hand, showed the smallest difference between Old and New subjects.

In the number of phonetic errors, the difference between Old and New subjects was large--the New subjects having the higher error rate for both consonants and vowels. The test results for the two groups are shown in Figure 2, computed for all presentations of each phoneme irrespective of its position in a syllable. (Note that the error scale in Figure 2 is ten times greater than in Figure 1.) The most striking proportional differences between the New and Old subjects' performances occurred with the consonants /ǰ/, /m/, /n/, /v/, and /w/, indicating that inexperienced listerers misinterpret these synthesized sounds but more experienced listeners can adapt themselves to some phonetic deficiencies. The vowel data showed that the groups' greatest differences lay in perceiving the phonemes /ʊ/, /ou/, /ɔi/, and /ɔ/, which were much more difficult for the New group.

When the syllable-initial and final consonant substitution data were separated and ranked, the error distributions took the form shown in Figure 3. Close similarities in the particular confusions made by Old and New subjects are evident. Both groups found /θ/ the most unintelligible phoneme in syllable-initial position (heard as /f/); /č/ and /š/ (both heard as /h/) and /t/ (heard as /k/) were the next most frequently misheard sounds. Error rates for all four voiceless phonemes exceed 15 percent. The errors themselves were the result of a misassignment of Place (to continue the use of articulatory terminology). The least confused phonemes for both groups were /f/, /s/, and /h/--all voiceless sounds--

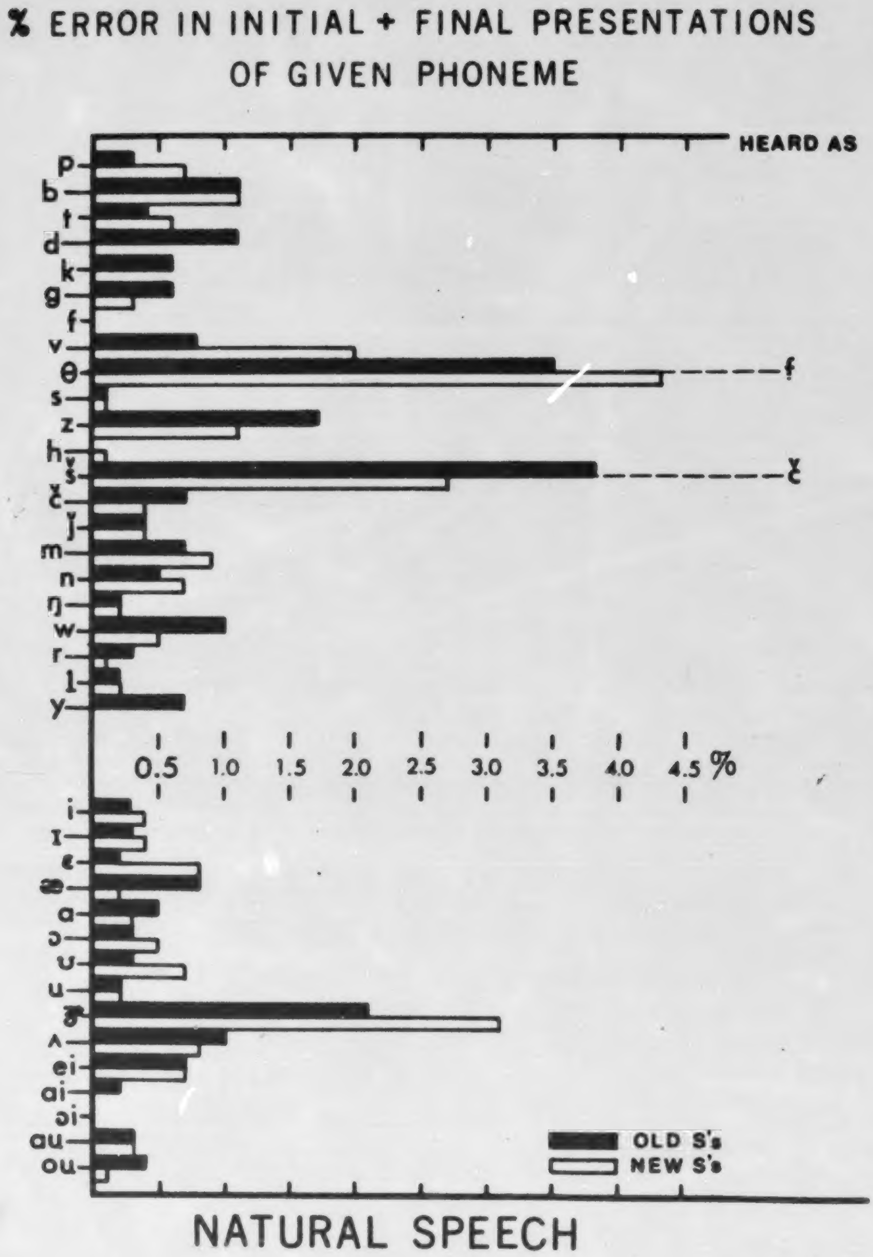
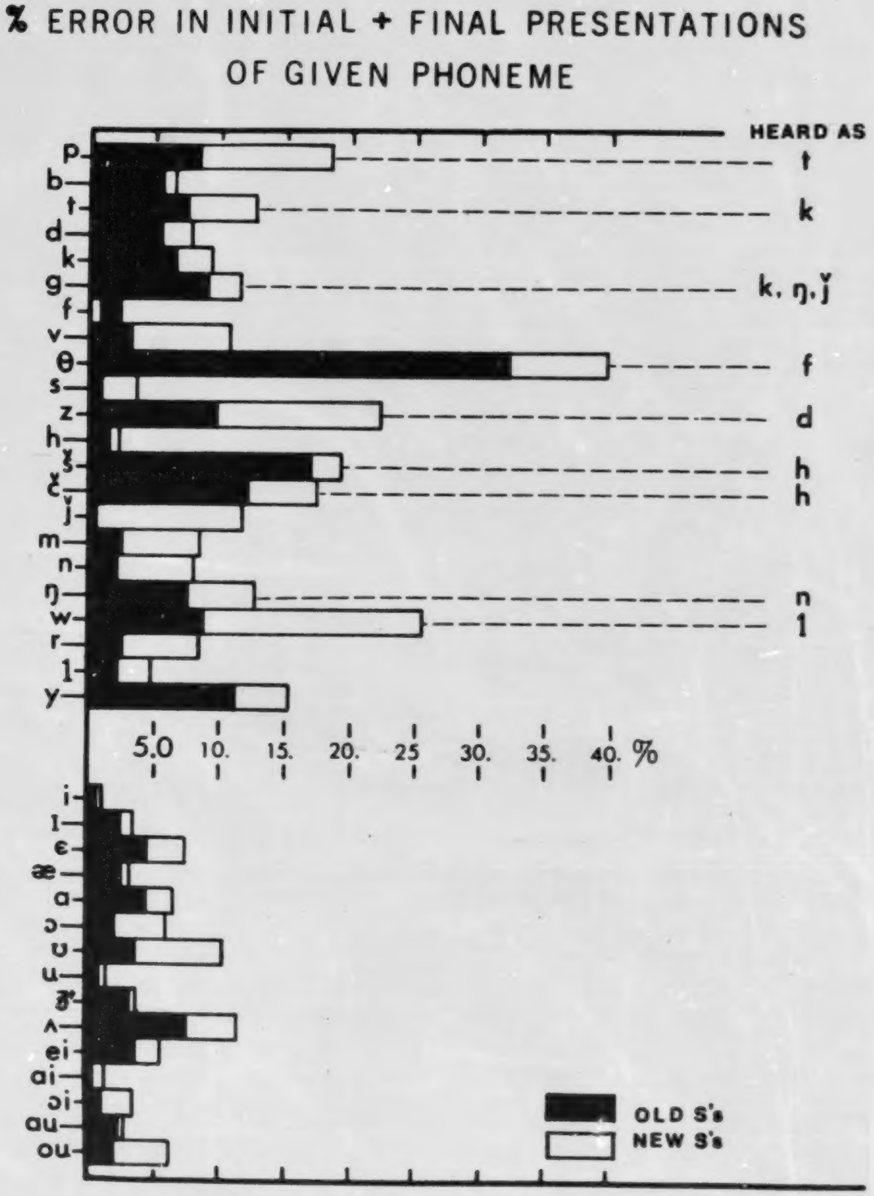

% ERROR IN INITIAL + FINAL PRESENTATIONS
OF GIVEN PHONEME
HEARD AS
p
b
t
d
k
g
f
v
θ
s
z
h
š
č
ǰ
m
n
ŋ
w
r
l
y
f
š
č
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5 %
i
ɪ
ɛ
æ
a
ɔ
ʊ
u
ɚ
ʌ
ei
ai
ɔi
au
ou
OLD S's
NEW S's
NATURAL SPEECH

FIGURE 1



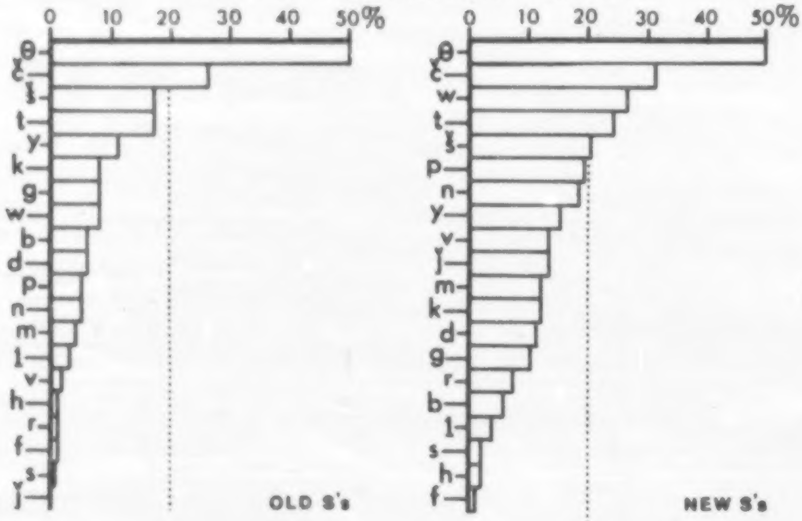
SYNTHETIC SPEECH

FIGURE 2

SUBSTITUTION ERRORS IN RANK ORDER IN SYNTHETIC SPEECH

(as percentage of times each phoneme was presented)

INITIAL CONSONANT ERRORS



FINAL CONSONANT ERRORS

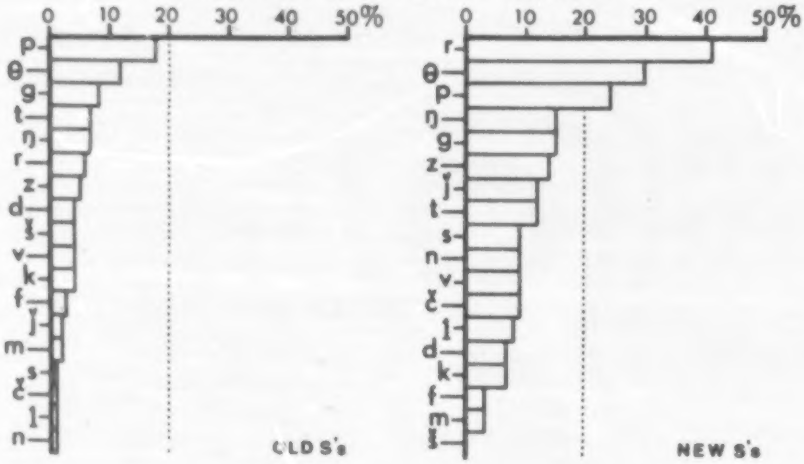


FIGURE 3

yielding error rates of less than 2 percent. In syllable-final position, the groups' responses continue to show agreement: the phonemes that produced most errors were /p/ (heard as /t/ or /k/), /g/ (heard as /k/), and /ŋ/ (heard as /n/ or /m/). Errors in /p/ and /ŋ/ are attributable to Place, and in /g/ to Voicing. Among the eight poorest final sounds there were more voiced than voiceless phonemes. The least difficulty (common to both groups of subjects) was with /m/, /f/, and /l/ in final position.

It is also noteworthy that /č/ and /š/ in initial position were very frequently misheard by both groups, but in final position were much more intelligible. The case of final /r/ was exceptional, with the New subjects hearing it as /l/ more than 40 percent of the time, although it produced only a 6 percent error rate for the Old subjects. Syllable position, in fact, was important in the perception of /r/ for both groups of subjects. In initial position, where /r/ should be consonantal, it produced few errors--but in final position, where /r/ should be vowel-like, it was the least intelligible phone for the New subjects and the sixth worst for the Old subjects. The nasals /n/ and /m/, in contrast, were more intelligible in syllable-final than in initial location.

Contextual Influences on Intelligibility

Various "coarticulation" effects at word boundaries are presumed to have contributed to the SNST overall error rate. The synthesis rules prescribe coarticulation (acoustically) taking a spectrum of contexts into account; nevertheless the varieties of potential clusters at and across the boundaries of abutting words is so large that rules have yet to be devised for many of the combinatorial possibilities. Thus, the intelligibility of "initial" and "final" vowels or consonants in context may be significantly lower than their intelligibility in isolation; this indeed proved to be so. Yet another property of the syntactically normal sentences, as presented, is their tendency to lead the listener astray. By unconsciously applying semantic constraints to the incoming sentence the listener may be caught off guard when an altogether "unexpected" word is heard. Consequently, the word may be misinterpreted or missed entirely.

Certain phonetic contexts were isolated as contributing to the high error rate for some phonemes. Heard as /f/, whether in initial or final syllable position (as had been the case in the natural speech), /θ/ produced greatest error when preceded by /n/ and when followed by /r/ or /ɔ/. Errors produced by an intended /č/ in syllable-initial position usually occurred after a voiced alveolar (/n/ or /d/) and before a high front vowel. In such cases /č/ was often heard as /h/, and occasionally as /k/. Poor in initial position only, /š/ gave most errors after /n/ and /d/ (i.e., the same context as the /č/ error), and tended to be heard as /h/--again like /č/--but /š/ was also heard as /f/, /s/, or /č/. When initial, /t/ was heard as /k/ in most cases of error, especially when it was followed immediately by /r/. The phoneme /t/ produced far fewer final substitutions than initial, but these few were heard as /d/ (after /i/, /ε/, /ɝ/, or /n/). Final /p/ errors were numerous when it was preceded by /i/, /ɪ/, or /ɑ/; /p/ being most often confused with /t/. Final /g/ error resulted in /ŋ/ or /ǰ/ substitutions (after /ε/), and /g/ was also heard as /k/ (after /ɑ/). Always syllable-final, /ŋ/ tended to be heard as /n/, and sometimes as /v/ or /m/, after /ɪ/ or /æ/. In summary, a preceding /n/ was a common environment for the misperception of /θ/, /č/, and /š/. Stops as a class, however, had no common environment found to produce consistent error.

Word location in the sentence was significant in the pattern of errors produced in the SNST, suggesting problems due to the stress rules or realization of the rules. Three levels of stress were used in synthesizing the test sentences. As pointed out earlier, under the speech synthesis rules currently in operation, the durations of the *mid* and *high* stress (equally long) are longer than the *low* stress, and the fundamental frequency of a *high*-stesssed syllable is raised slightly above that of *mid* (or *low*) stress. For the Old and New subjects pooled, the natural speech produced most errors on the third response word (the verb) of the sentences, but the greatest number of synthetic speech errors consistently occurred on the second word, the noun "subject" (see Figures 4a and 4b). Examination of the Omission (i.e., "No Response") and Substitution errors in Figure 5 and Figures 6a and 6b shows a similar trend. The reason for this trend is not yet clear, but it may originate in the interaction of memory load and the amount of extra attention that must be devoted to interpreting the synthetic speech sounds; the stress assignment or realization may also be involved.

Comparison of SNST and MRT Results

Several differences between findings with the MRT and SNST should be noted. The natural speech version of the MRT (a closed-response test employing isolated words) produced the classic result that initial consonants were more intelligible than final consonants. The SNST, however, indicated that the intelligibility of natural speech words was essentially equal in the two positions. In synthetic speech the result was different again: in both the MRT and SNST tests, the word-final consonants proved to be somewhat more intelligible than initial consonants. The explanation for this oddity remains to be determined; it clearly concerns the synthesis rules and/or the synthesizer itself.

There is also some agreement between the results of the two types of test (MRT and SNST). The highly intelligible phonemes in the SNST synthetic speech data of both groups of subjects for both initial and final positions were /f/, /s/, and /l/. The most intelligible phonemes in the MRT for both syllable positions were /*f*/, /t/, /*s*/, and /g/ (phonemes in common to SNST and MRT are underlined). In final position in the syllable, the least intelligible phones were also similar in the two tests--in the SNST, /θ/, /p/, /r/, /ŋ/, and /g/, and in the MRT, /*r*/, /b/, /*θ*/, and /*ŋ*/. (Note that /b/ was not presented finally in the SNST.) In initial position, however, there was no agreement on the least intelligible sounds in the two tests--in the SNST, /θ/, /č/, /t/, and /š/ were the poorest phonemes; in the MRT, /v/, /n/, and /h/ were the worst. (Note again that /θ/ did not appear in the MRT in initial position.)

Despite these initial position differences, both the closed- and the open-response tests have demonstrated that certain synthetic speech sounds are much more intelligible in one syllable position than another. Both tests also have shown that Place errors dominate in either syllable position, but that nasal phonemes tend to produce Manner errors in initial position. Put another way, the phonemes of the synthetic speech tested are usually confused with phonemes within the same Manner and Voicing class, except for initial nasals which are likely to be heard as nonnasals (i.e., as stops or as semivowels).

Learning Effects in Synthetic Speech

Although further testing of the learning effects observed with synthetic speech in the MRT was not a prime objective of the SNST, the fact that the New

WORD LOCATIONS IN WHICH SUBSTITUTION ERRORS OCCURRED

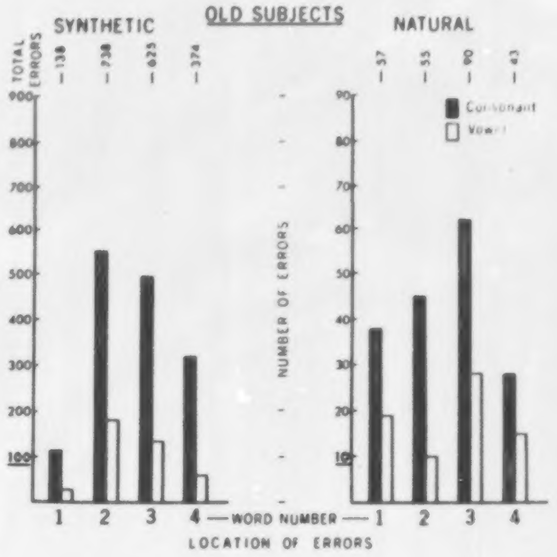


FIGURE 4A

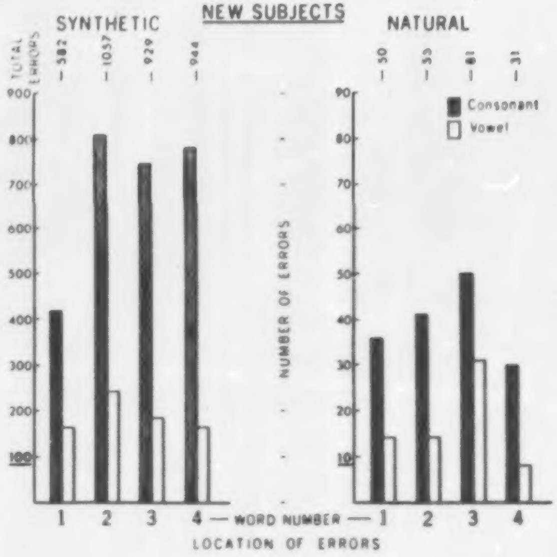


FIGURE 4B

WORD LOCATIONS IN WHICH "NO RESPONSE" ERRORS OCCURRED

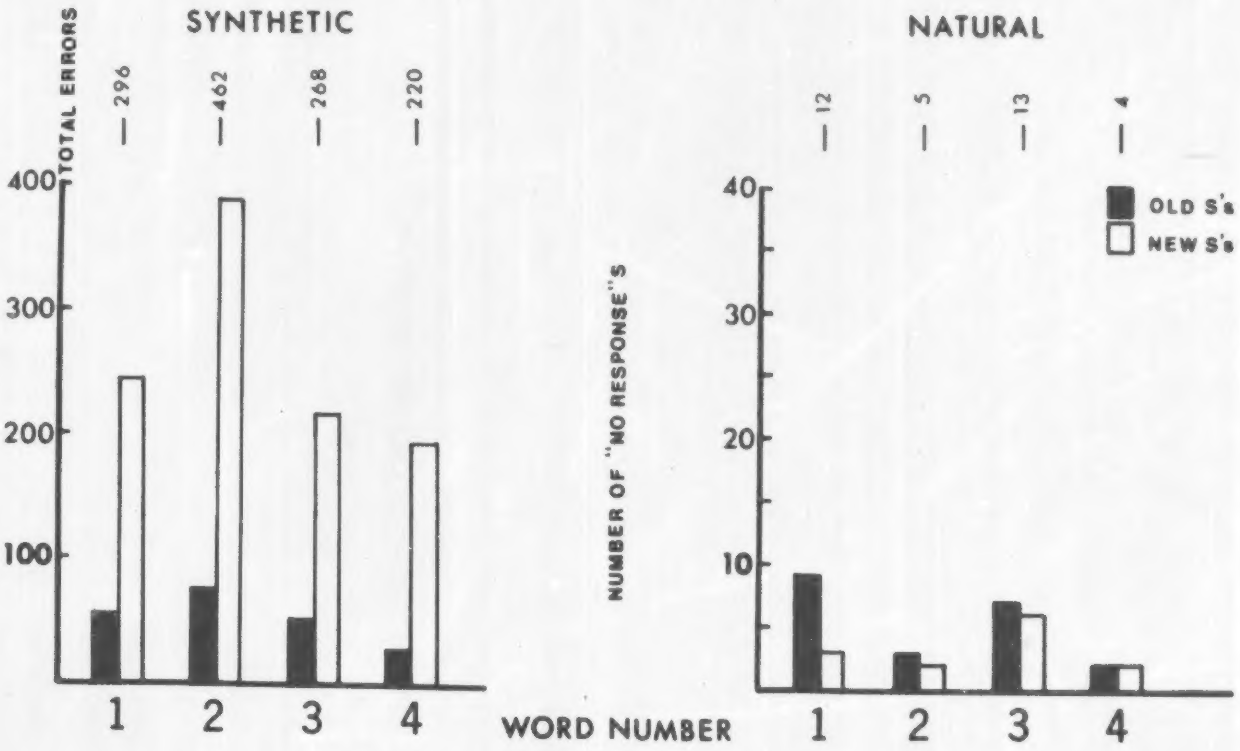


FIGURE 5A

FIGURE 5B

WORD LOCATIONS IN WHICH SUBSTITION ERRORS OCCURRED

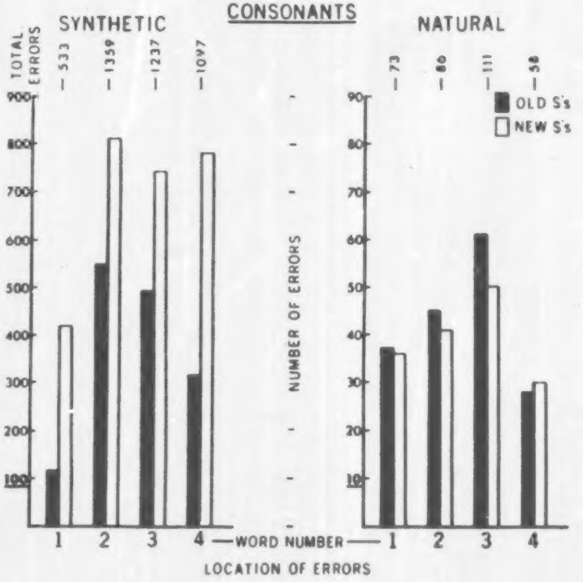


FIGURE 6A

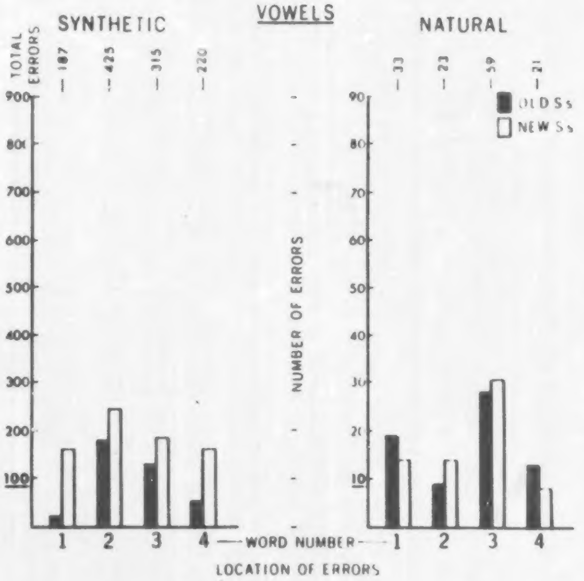


FIGURE 6B

subjects produced significantly more errors than the Old subjects in the phonetic task, and not in the memory task, is noteworthy. A comparison of the most intelligible phonemes for the Old group versus the New group (Figures 7a and 7b) shows that in common, the two groups found /m^2, f, s^1, l^1, š^2, h/* highly intelligible; but that in addition, the Old subjects also found /p^1, m^1, w, v, d^2, s^2, z^2, n, l^2, r^1, č^2, ǰ, k^2/ among the highly intelligible.

And, in comparison to the MRT results, the Old subjects did better in the SNST in identifying /p, m, w, v, n, l, ǰ/ in syllable-initial position, and /v, d, l, š/ in syllable-final position. In the SNST, however, /t, d^1, g/ were less intelligible than they had been in the earlier test. Fricative, nasal, and glide/liquid perception thus was improved in the SNST, although, on balance, stop perception was not improved. (The hardware synthesizer and the rules for synthesis were identical in the two tests.)

CONCLUSIONS

Summary

The SNST test was composed of monosyllabic, high-frequency English content words in short sentences that were syntactically normal but semantically anomalous. The words contained nearly all of the English phonemes, located in word-initial and word-final position, and reflected the relative frequencies of phones in the language as a whole. The intent of the test was to assess the intelligibility of the phonemes in context, and to discover the substitutions made for the least intelligible phonemes--in order to get information that might lead to improvements in the Haskins Laboratories' synthesis program. A parallel test in natural speech was run to obtain a standard against which the intelligibility of the synthesized speech might be compared. The results reported here reflect the performance of a specific synthesis program in combination with one hardware synthesizer, and therefore are of particular relevance to the designers and users of the Haskins Laboratories' hardware and software. However, the form of the test itself, "meaningless" sentences containing normal grammatical sequences but abnormal semantic combinations, is an approach to intelligibility testing that has not been widely exploited. This technique, involving connected speech stimuli and an open mode of response, has revealed clear differences in the location of errors (syllable-initial versus final, and word position in sentence) between natural speech and synthetic speech. Further, the SNST has highlighted the degree of difficulty in perceiving synthetic phones in monosyllables much more sharply than did the Modified Rhyme Test.

As was also found previously with the MRT, the overall intelligibility of final consonants in synthetic speech is better than that of initial consonants. The SNST natural speech control test produced a very low error rate for both syllable positions. The individual phonemes producing errors, however, differed in the two positions both in natural speech and in synthetic speech.

---

*In the MRT, the most intelligible phones were /m^2, f, t, d^1, s, z^2, n^2, r^1, č^2, ǰ^2, k^2, g/. (Superscripts 1 and 2 denote syllable-initial position and syllable-final position, respectively.)

SYNTHETIC SPEECH: PHONEMES WITH HIGHEST INTELLIGIBILITY

OLD SUBJECTS

	LABIAL	LAB-DENT	DENTAL	ALVEOLAR	PALATAL	VELAR	GLOTTAL
STOP Voiceless	p^1					k^2	
STOP Voiced				d^2			
FRICATIVE Voiceless		f		s	$\check{s}^2, \check{c}^2$		h
FRICATIVE Voiced		v		z^{*2}	$\check{j}$		
NASAL	m			n			
LIQUID	w			ℓ, r^1			

FIGURE 7A

NEW SUBJECTS

L	L -D	D	A	P	V	G
	f		s^1	$\check{s}^2$		h
m^2						
			ℓ^1			

FIGURE 7B

LEGEND
Phoneme presented in final position only: *
Phoneme good in initial position only: ^{1}
Phoneme good in final position only: ^{2}
(/ð/ and /ž/ did not appear in the test.)

Limitations of the SNST

Like the MRT which preceded it, the SNST had its own shortcomings. Memory problems surely compounded the errors made in phoneme identification; the anomalous semantic content of each sentence may have confounded the listeners who inadvertantly attended to the word sequences as if they were meaningful; and the distribution of abutting phonemes across word boundaries was not controlled and probably led to sporadic interword effects. Further, consonantal clusters were occasionally included in the stimuli (beclouding the issue of individual phoneme confusability) because the high-frequency monosyllables that fit the SNST grammatical requirements were entered into the stimulus supply without careful phonetic analysis. Finally, the particular stress and intonation rules and their realizations in synthesis may well have contributed to the unintelligibility of words in certain sentence locations. The disparity between the findings in natural speech and those in synthetic speech in respect to the word location of the greatest substitution (and other) errors suggests that either stress assignment rules or stress realization rules--or both--contribute to the error. (The third word in a sentence produced the highest error in natural speech, but the second held that distinction in synthetic speech.)

Although it employed a wider repertoire of phonemes than the MRT, the SNST nevertheless had its distributional limitations. Due to the structural repertoire of the test (containing CVC monosyllabic nouns, adjectives, and verbs) the phonemes /ð/ and /ž/ were never presented in the test words. Also, because the inventory of words was drawn from English and reflected its phoneme distribution and frequency, /ž/ was not used in word-initial position, nor was /b/ used in final position. Of course /h/ and /ŋ/ occurred only in initial and final position, respectively. [The semivowels /w/ and /y/ were counted among the consonants when used in syllable-initial position, but were subsumed within the vowel classification (as dipthongs) when they occurred in final position.] Thus the test necessarily omitted two phonemes entirely (/ð/ and /z/) and omitted one sound initially (/z/) and one finally (/b/). Table 3 lists the SNST's total number of phoneme presentations to each group (of 16 listeners).

Final Comments and Future Plans

The design of the SNST provided the listeners with utterance structures and response choices that were more normal than those in the earlier MRT. The substitution errors made in the SNST can be reasonably accepted as sounds actually heard by the listeners (insofar as there were existing English words needed by the naive listeners to use as their responses). A fair degree of consistency was found in the syllable environments in which the major substitution errors occurred, indicating specific frames for allophone rule improvement. However, the syllable boundary contexts were not controlled in the SNST (e.g., final voiceless stops were not held constant before syllable-initial nasals, and vice versa). This facet of synthetic speech will have to be examined intensively in future testing.

The prosodic properties of the stimuli, and the context, must also be further controlled as factors in phoneme and word intelligibility. Spectrograms made of ten of the test sentences in both the synthetic and the natural speech versions showed fairly good syllable duration agreement in both sets. (It will be recalled that the human reader attempted to reproduce the timing and inflection of each synthetic sentence immediately after hearing it.) Nonetheless, an examination of the spectrograms shows that for the human talker, the word "the"

TABLE 3

TOTAL OCCURRENCES OF PHONEMES PRESENTED TO EACH GROUP OF 16 LISTENERS

(Initial and Final Syllable Positions Summed)

CONSONANTS		VOWELS	
r	3584	ɛ	1632
t	3472	eɪ	1376
d	3376	æ	1312
l	2816	ɔ	1280
s	2592	i	1136
n	2160	ɪ	1008
k	1920	ɑ	992
b	1584	ou	944
p	1184	ɑi	864
g	1040	ʌ	592
f	944	u	528
h	944	ɑu	336
m	758	ʊ	288
š	640	ɝ	288
w	608	ɔi	208
ŋ	464		
č	448		
v	256		
θ	256		
ǰ	224		
z	176		
y	144		
(ð)	NOT PRESENTED AS STIMULUS		
(ž)	" " " " (Impossible as word initial, very rare as final)		

(which occurred twice in each sentence) was markedly shorter, and lower in amplitude, than it was in the synthetic versions; and the third word--the verb--of each sentence was generally shorter for the human than for the synthesizer.

Rhythmic effects are apparent in real speech (e.g., syllable durations and other prosodic parameters are conditioned by contextual stress) but the study of speech rhythm in prose is young. Therefore the synthesis rules used in the SNST stimuli understandably fail to take interword rhythm as such into account. This area, in conjunction with other aspects of utterance prosody, is under investigation.

REFERENCES

Nye, P. W. and J. H. Gaitenby. (1973) Consonant intelligibility in synthetic speech and in a natural speech control (Modified Rhyme Test results). Haskins Laboratories Status Report on Speech Research SR-33, 77-91.

Thorndike, E. L. and I. Lorge. (1968) The Teacher's Word Book of 30,000 Words. (New York: Teachers College Press).

A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech*

Paul Mermelstein
Haskins Laboratories, New Haven, Conn.

This paper considers a sequential strategy for acoustic-phonetic speech analysis. Each analysis process is applied to an appropriately labeled speech segment and results in a possible subsegmentation of the original segment. The segments resulting from the analysis are labeled according to the analysis results.

The advantages of the strategy are that no more segments are considered than those actually differentiated by the analysis steps. The extraction of acoustic cues pertinent to a phonetic feature can be tuned to classes of sounds separated on the basis of other cues and this serves to increase the reliability of segment labeling. The analysis sequence yields a structure for the syllabic units of the speech signal that may be used to retrieve similar syllabic units for detailed comparison.

Introduction

What is the relationship between the acoustic cues of the speech signal and its phonetic features? Evidence available today appears to indicate that there is no simple transformation from the cues directly extractable from the signal by signal processing techniques to the phonetic features, the distinguishing characteristics of the individual phonetic elements. Rather, a complex encoding takes place so that information about a particular feature of a segment may in fact be carried by neighboring segments. A feature may be signaled by cues that differ depending on other features present in the same segment, as well as on the contextual environment in which that segment is embedded. This paper outlines a strategy for drawing inferences about the phonetic features of segments from a sequence of acoustic processing steps, each of which characterizes in increasing detail the acoustic information present.

---

*Presented at the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Pa., 15-19 April 1974; to be published in the symposium proceedings.

Acknowledgment: The author appreciates the assistance of F. S. Cooper, G. Kuhn, and L. Lisker, who participated in discussions leading to the formulation of the ideas presented here.

A syllabic unit is defined as a segment of the speech signal delimited by significant minima in a loudness function, a time-smoothed frequency-weighted summation of the signal spectrum. We focus our attention on characterizing segments within a syllabic unit and on the relationships between these segments. The relationships between the segments reveal a structure for the syllabic unit which may be used to select units of similar structure from a store of syllabic forms. At any point within the processing sequence the cost of further characterization of the segments is weighed against the remaining ambiguities in the possible structural matches. The sequence of acoustic analysis steps set up as generally applicable may be modified, in the light of the ambiguities remaining at any point, to derive the maximum useful information from any particular analysis step.

Strategy for Segmental Analysis

Following Fant (1962), I consider the speech signal to be composed of "a sequence of minimal sound segments, the boundaries of which are defined by relative distinct changes in the speech wave structure" (p. 7). Consider a sequence of one or more such minimal sound segments that have some common acoustic property as an acoustic segment. By focusing in turn on different properties, we can isolate and appropriately label segments that exhibit these properties to differing extents. We may contrast two strategies for the segmentation and phonetic labeling of acoustic segments:

a) A set of acoustic cue detectors is constructed to operate on the speech signal in parallel and independently of each other. Whenever a change is noted in at least one acoustic cue, the signal is divided into separate segments. Call this the parallel cue detection strategy.

b) A number of acoustic cue detectors are applied to the speech stream sequentially. The selection of the detector to be applied next follows a decision-tree. Call this the sequential cue detection strategy.

The parallel detection strategy is applicable to a model of speech analysis that considers the momentary speech signal to be a function, probably nonlinear, of independent acoustic features. Certain features manifest themselves quite independently. Voicing and frication can be considered independent features from this point of view. The amount of aperiodic energy needed to call a segment fricative in the presence of voicing is larger than that required to call it fricative in the absence of voicing. In fact, the outputs due to the separate excitation sources are known to combine nonlinearly. All sound segments are searched for all features. As a result, the sound segment is located in the hyperspace of acoustic features.

The sequential strategy makes use only of a minimal set of cues adequate to characterize the sound segment. Cues outside the minimal set are considered redundant. Since the phonological units are not always represented by the same acoustic cues, acoustic properties considered redundant in some cases may be used to aid the general transformation from acoustic cues to phonetic segments. There exists evidence today for independent human storage of phonetic features, for example, place and manner of production (Wickelgren, 1966). The corresponding acoustic cues are, however, not generally independent. The perception of an acoustic cue underlying a particular phonetic feature may vary with changes in the acoustic cues underlying other phonetic features (Pisoni and Sawusch, 1973). For example, place of production cues are functions of the voicing feature, although the voicing cues are generally independent of the place feature. Therefore, independent search for acoustic cues appears undesirable.

An important property of the sequential strategy is that segmentation and labeling are results of the same operations. A stretch of the speech signal is segmented if some analysis operation yields significantly different results over that stretch. Simultaneously, differing labels are attached to the newly derived segments. This procedure runs counter to the traditional pattern recognition strategy of complete segmentation followed by analysis of segments.

Application of any particular analysis function to an appropriate acoustic segment can yield only a small number of alternative productions, as suggested by the phonological rules of the language. By thus limiting the number of segments produced we avoid the requirement for independent analysis of time-synchronous chunks of speech, we limit the total number of decisions made, and we reduce the possibilities for phonologically inconsistent labeling of segments.

Analysis Rules

The segmentation and analysis rewrite rules given below are formally context independent. The decision whether a segment is to be further subdivided is based only on acoustic information contained in that segment. These rules govern mainly the number of segments to be produced and their labeling as to voicing and manner of production. Further place of production analysis rules can be expected to be context dependent and they will be considered in greater detail below. The entire strategy consists of two stages of analysis, one context independent, determining the number of subsegments, and one context dependent, deriving further information about the individual segments.

In constructing an appropriate acoustic analysis sequence, we may profitably utilize the phonological rules of the language that restrict the segmental makeup of syllabic units. The rules on the manner of production of the sequence of units are particularly strong. For example, a syllabic unit must be completely voiced, be composed of a voiced segment bounded by voiceless segments on one or both sides, or be a syllabic fragment that is completely unvoiced. Segments differing in voicing and manner of production can be ordered according to sonority so that in going from the initial boundary through the syllabic peak to the final boundary, sonority is first monotonically increasing then monotonically decreasing. This suggests that a strategy for manner of production cue analysis proceed from the edges of the syllabic unit to the center and look for manner of production cues that accompany increasing degrees of sonority.

The place of production rules, not yet implemented, allow the analysis of segments to be carried out according to a sequence dependent on the previously derived manner of production information. The selection of vowels appears to be least dependent on the neighboring consonantal segments; therefore a preliminary vowel decision can be made first. This preliminary decision, classifying the vowels only into three groups, /i/-, /a/-, and /u/-like, can be followed by a subsequent analysis taking coarticulation rules into account once the neighboring consonants have been identified in greater detail. Determination of the place of production of consonant classes is context dependent in the sense that the syllabic vowel color is taken into consideration when making that decision. Consonants are considered in order of decreasing sonority moving outward from the syllabic peak.

Implementation of Rewrite Rules

The following rewrite rules for segmentation and labeling have been implemented and are undergoing evaluation. Each rule transforms the given segment into the indicated subsegments if the criteria for the results of the acoustic analysis are satisfied. The subsegments correspond to the nodes on the segment-structure tree that are descendants of the original segment.

Note that some intervocalic segments will be cut apart by the syllabic division rule. Thereafter the two subsegments will be processed individually as parts of each syllabic unit. When the final results are analyzed, identically labeled segments that follow each other in time may be combined into one segment. The following nine rules give the voicing and manner of production analysis (Table 1). Further rules are to be added for analysis of vowels and place of production of consonants.

Our syllabic units are acoustic segments. They are derived from the actual production and thus do not correspond precisely to linguistic (phonological) syllables. In particular, two words may form one syllabic unit if the first ends in an open vowel, the second starts with an open vowel, and no glottal stop intervenes. For example, "the old" generally forms one syllabic unit /ðIold/ in which rule 9 will attempt to find two vowels. Rules 7 and 8 are separated to indicate explicitly that the cues for prevocalic liquids may be different from those for postvocalic liquids.

The rewrite rules cited are not meant to be complete. Rather they indicate the kinds of rules required to implement the strategy outlined here.

Data Structures

The question of an appropriate data structure to express the results of the analysis operations is rather important. The structure most appropriate for hierarchic analysis is that of a tree whose root node corresponds to the complete utterance and whose subtrees correspond to each syllabic unit. Where an acoustic property is found present or absent for the entire segment, that segment is relabeled but not cut. Where a significant change is found for that property over the span of the segment, the segment is cut into two parts at the point of change. Branches equal in number to the number of subsegments are grown from the node corresponding to the original segment. The nodes branching from any higher node are ordered in time according to the time ordering of the segments corresponding to the nodes. Since the root nodes corresponding to the syllabic units are similarly ordered, a temporal chain of segments is maintained at all times in the processing sequence. This chain allows efficient reference to preceding and succeeding segments (nodes) even where these do not branch from the same parent node. By ordering the analysis operations so that those operations which can be expected to be more reliable are carried out first, we can construct a metric of differences for syllabic units that decreases with the level of the node where a difference is encountered. Highly similar units have the same structure and differ only in the label assigned to the terminal nodes. Less similar units may

TABLE 1

1. $$[\text{Sentence}] \rightarrow [\text{Syllabic unit}]\ (\text{Sentence})$$ //Syllabication//

2. $$[\text{Syllabic unit}] \rightarrow \left\{\begin{array}{l}[\text{Voiceless segment}]\\ (\text{Voiceless segment})\ [V_1]\ (\text{Voiceless segment})\end{array}\right\}\ (\text{Silent segment})$$ //Voicing decision//

3. $$[\text{Voiceless segment}] \rightarrow \left\{\begin{array}{c}(\text{Voiceless burst})\ \left\{\begin{array}{l}[\text{Voiceless fricative}]\\ [\text{Aspiration}]\end{array}\right\}\\ [\text{Voiceless burst}]\end{array}\right\}$$ //Voiceless subsegments//

4. $$[V_1] \rightarrow (\text{Voice bar})\ (\text{Voiced burst})\ [V_2]\ (\text{Voice bar})$$ //Voiced stops//

5. $$[V_2] \rightarrow (\text{Voiced fricative})\ [V_3]\ (\text{Voiced fricative})$$ //Voiced fricatives//

6. $$[V_3] \rightarrow \left\{\begin{array}{c}(\text{Nasal consonant})\ [V_4]\ (\text{Nasal consonant})\\ [\text{Syllabic nasal}]\end{array}\right\}$$ //Nasal//

7. $$[V_4] \rightarrow (\text{Liquid})\ [V_5]$$ //Prevocalic liquid//

8. $$[V_5] \rightarrow \left\{\begin{array}{l}[V_6]\ (\text{Liquid})\\ [\text{Syllabic liquid}]\end{array}\right\}$$ //Postvocalic liquid//

9. $$[V_6] \rightarrow \left\{\begin{array}{c}\left\{\begin{array}{l}(\text{Short vowel})\\ (\text{Semivowel})\end{array}\right\}\ [\text{Long vowel}]\\ [\text{Long vowel}]\ \left\{\begin{array}{l}(\text{Semivowel})\\ (\text{Short vowel})\end{array}\right\}\\ (\text{Short vowel})\end{array}\right\}$$ //Vowel-like segments//

Legend:
[ ] - mandatory segment
( ) - optional segment
{ } - ordered disjunction
// // - comments

FIGURE 1

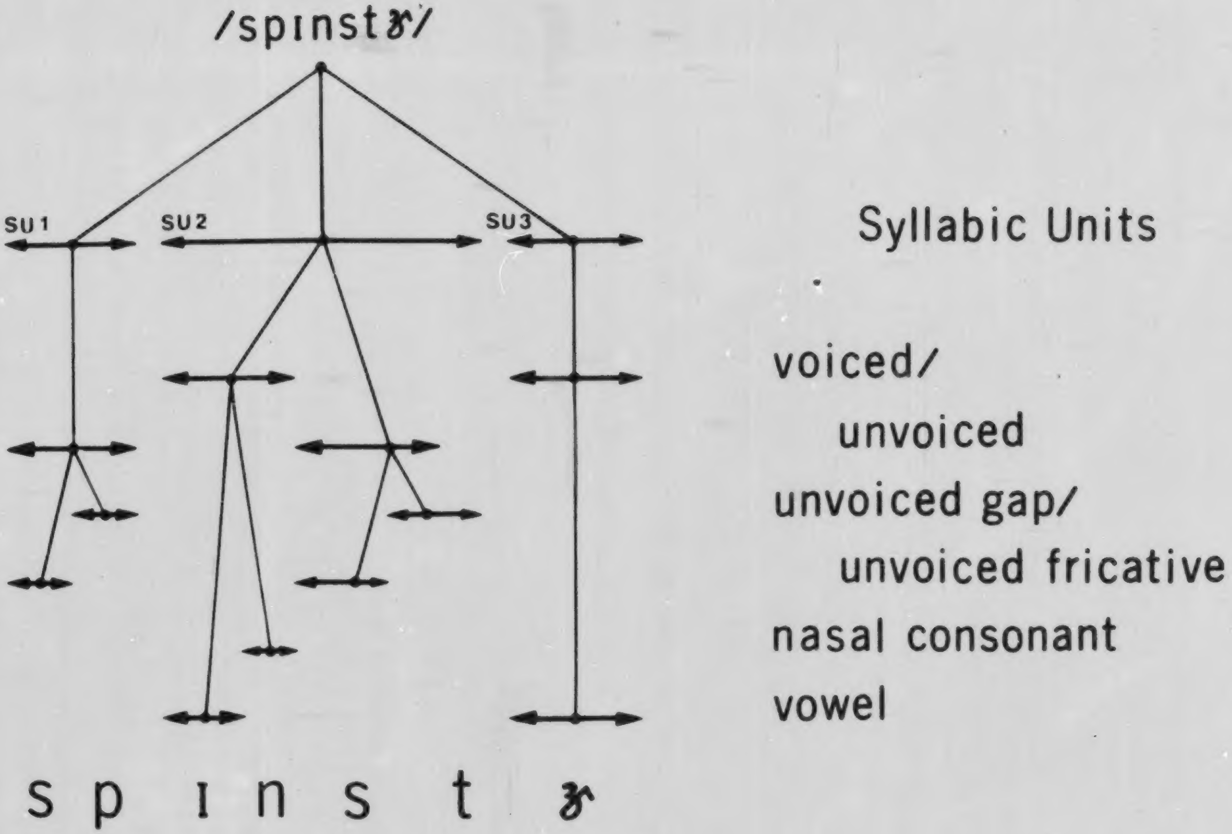

/spɪnstɚ/
SU1
SU2
SU3
Syllabic Units
voiced/
unvoiced
unvoiced gap/
unvoiced fricative
nasal consonant
vowel
s p ɪ n s t ɚ

Figure 1: Segment tree for word "spinster."

have a common set of topmost nodes to which differing subtrees appear to be attached.

We may assign a similarly structured map to each entry in the lexicon of admissible syllabic forms. Consider the situation where a phonologically acceptable derivation is found, but an error is suspected because no entry matching the derived structure can be located in the lexicon. Likely substitutions can be found by examining forms to which only a partial match exists. To this end, the lexical entries are assigned a code appropriate to the structural map. Retrieval is thus made possible on the basis of the complete code word or any initial subsequence thereof.

Figure 1 illustrates the segmentation-derived structure for the word "spinster." Three syllabic units are found, the first a fragment containing the initial fricative [s] and the stop gap of [p], the second roughly corresponding to the sequence [inst], and the third the final vowel [ɚ]. The voicing decision separates the second syllabic unit into two parts, finds the first all unvoiced and the last all voiced. The unvoiced fricative detector segments the first syllabic unit and the unvoiced segment of the second. The nasal detector is applied only to voiced segments and it segments that portion of the second syllabic unit. The fact that seven segments are found for this word, equal in number to the constituent phonemes, is purely coincidental. The total number of segments may be larger than the number of phonemes, due to segmentation of stop releases, or smaller, due to incomplete segmentation of vowel-like sequences.

To date, the voicing, syllabic unit, frication, and nasal consonant indicators have been implemented. These are the distinctly different segments of the speech stream. The detection of liquids and semivowels lies immediately ahead. Thereafter attention will focus on the information supplied by segments within the syllabic unit regarding the place of production of those segments themselves as well as of neighboring segments. Although to date our experiments have been restricted to short utterances, there appear few serious problems in extending the procedure to longer utterances. Perhaps the major problem apparent at the moment is the increased difficulty of segmenting unstressed syllabic units and the decrease in detail recoverable from them. This may require the treatment of the stressed and unstressed counterparts of the same syllabic units as different lexical forms for comparison purposes.

REFERENCES

Fant, C. G. M. (1962) Descriptive analysis of the acoustics of speech. Logos 5, 3-17.

Pisoni, D. B. and J. R. Sawusch. (1973) On the identification of place and voicing features in synthetic stop consonants. Haskins Laboratories Status Report on Speech Research SR-35/36, 65-80.

Wickelgren, W. A. (1966) Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Amer. 39, 388-398.

What Information Enables a Listener to Map a Talker's Vowel Space?*

Robert Verbrugge,+ Winifred Strange,+ and Donald Shankweiler++

The acoustic structure of vowels varies markedly from one speaker to another, and from one phonemic context to another (Peterson and Barney, 1952; Peterson, 1961; Stevens and House, 1963). It is commonly assumed that a listener identifies a talker's vowels in terms of the relation between their acoustic structure and the acoustic structure of other vowels produced by the same person (Joos, 1948; Ladefoged and Broadbent, 1957; Ladefoged, 1967). It is also common to speak of vowels as situated in a vowel space, the shape of which is a function of each individual's vocal tract characteristics. As a result, a talker's vowel space would be completely specified only after a listener hears an extended sample of the talker's speech. Such experience would allow the listener to calibrate or normalize to each particular voice he encounters. This suggests that the largest source of errors in identifying vowels will be inadequate exposure to a novel voice, i.e., hearing an utterance that is too brief or impoverished to allow accurate calibration.

Our first study was an attempt to assess this claim quantitatively, by comparing the identification of vowels under two conditions. In the Mixed Condition a large number of talkers spoke a series of syllables; on any one syllable the listener encountered a voice that was unfamiliar and unpredictable. In the Blocked Condition subjects heard the same series of syllables spoken by one person, so there was ample opportunity to become familiar with the voice, and the talker was fully predictable from one syllable to the next.

---

*Paper presented at the 87th meeting of the Acoustical Society of America, New York, 25 April 1974.

+University of Minnesota, Minneapolis.

++Haskins Laboratories, New Haven, Conn., and University of Connecticut, Storrs.

Acknowledgment: This paper and the following one report a portion of research begun during the academic year 1972-73 while D. Shankweiler was a guest investigator at the Center for Research in Human Learning, University of Minnesota, Minneapolis. The work was supported in part by a grant from the National Institute of Child Health and Human Development to the Center, and in part by grants awarded to Shankweiler and to J. J. Jenkins by the National Institute of Mental Health. We wish to thank Thomas Edman and Kevin Jones for their assistance in every phase of the experimental work and James Jenkins for his advice and encouragement.

Nine vowels appeared in a fixed consonant frame, /p-p/, to form the following syllables: /pip, pɪp, pɛp, pæp, pɑp, pɔp, pʌp, pʊp, pup/. Each of the nine syllables was spoken five times, for a total of 45 tokens per test.

In the Mixed Condition, 15 talkers were chosen: 5 men, 5 women, and 5 children, representing a wide variety of vocal tract sizes and fundamental frequencies. Each of the 15 people spoke three different vowels during the test.[1] The three tokens for each talker were separated by at least eight other talkers. Listeners heard the test twice, making a total of 90 judgments, 10 for each vowel. They recorded their judgments by circling the appropriate word on an answer sheet.

In the Blocked Condition, a representative man, woman, and child each spoke the full series of 45 test items. Listeners heard each of the three tapes, in one of several orders. Data for only the first two repetitions were pooled together across groups to keep the scores comparable to the Mixed Condition, i.e., five judgments per vowel in a first repetition and five judgments in a second repetition. A judgment was considered an "error" if the indicated vowel was placed in a different phonemic category than that intended by the experimenters. The error measure,[2] then, is a compound of talker and listener processing. In this account of our experiments, we do not attempt to separate these sources.

Listeners made an average of 17 percent errors in identifying vowels produced by the panel of randomly ordered talkers (the Mixed Condition), while in the Blocked Condition, listeners averaged 9.3 percent errors for the vowels of the three single talkers. Thus, it is plain that familiarity with a talker's voice significantly improved the accuracy of identification, though less than half of the errors can be attributed to this source.

There are two ways to look at these error percentages. First, 9 percent is a relatively high "error" rate, considering the complete predictability from trial to trial of both the speaker's voice and the consonantal frame; there are sources of vowel ambiguity not attributable to uncertainties in calibration. Second, 17 percent is a relatively low error rate, given that each judgment is made without any prior experience with the voice and without the benefit of sentential context. Clearly there is a great deal of information within a single syllable which specifies the identity of its vowel nucleus. [Peterson and Barney (1952) report an even lower error rate, 5.6 percent, for 10 vowels in /h-d/ context with 10 talkers randomly mixed on each test.]

These data challenge the assumption that extended familiarization with a vowel space is the primary factor controlling vowel identification. The question of what *are* the primary contributors must be reopened for study. In the following paper we report our studies of the information available within a single syllable. In the present paper we consider phonetic information that may extend across several syllables.

---

[1] The talkers read the test syllables which were printed individually on cards. Standard English orthography served to represent seven of the syllables. In the two cases in which the target syllable was not a word, /pɔp/ and /pʊp/, the card specified "vowel as in cawed," and "vowel as in could." Most talkers pronounced the target syllables without hesitation and in most cases their tokens were recorded on magnetic tape after a single rehearsal. In no case did the experimenters provide a talker with spoken models.

[2] Failures to respond were counted as errors. These occasions represent less than 2 percent of the total errors.

Because listeners' identification of vowels was better in the Blocked Condition than in the Mixed Condition, we may infer that information specifying the vowel must have been carried over the series of utterances of a single talker. We will first examine the advantage of keeping the talker constant on a vowel-by-vowel basis. Then we will test one hypothesis about the source of the information conveyed by talker constancy.

The errors made in identifying each intended vowel are shown in Figure 1; the columns indicate the percent of the time listeners made errors on each of the nine vowels. The hatched columns represent percent errors in the Mixed Condition, while the white columns represent errors in the Blocked Condition. For almost every vowel the percentage of errors drops in the Blocked Condition; the only exceptions are for /ɪ/ and /ɑ/. Three vowels, /i, ɪ, u/, are readily identified in either condition. Of the remaining six vowels which are relatively ambiguous, only /ɑ/ fails to show improvement, while familiarization definitely aids perception of /ɛ, æ, ɔ, ʌ, ʊ/.

But can we be sure that the improvements we observe are genuine? Shifts in response biases from one condition to the next could be responsible for some of these apparent improvements. A vowel could be correctly identified more often simply because it is more popular as a response. A direct sign of such a response bias is how often the vowel is used as an incorrect response to other vowels; when the vowel becomes more popular, the frequency of these false identifications increases.

The horizontal axis in Figure 2 indicates the change in correct identification between the Blocked and Mixed Conditions; placement to the right of the central vertical line represents superior performance on the Blocked Condition over that on the Mixed Condition. The vertical axis indicates the change in false identifications; placement above the central horizontal line represents greater frequency of false identifications on the Blocked Condition relative to the Mixed Condition.

True improvement may be defined as an increase in correct responses, coupled with a decrease in false identifications. Four of the more ambiguous vowels, /ɛ, æ, ʌ, ʊ/, show genuine improvement by this measure. On the other hand, a change in correct identification that correlates with a change in false identification may be attributed to response biases alone. Thus the apparent improvement for /ɔ/ may be attributed to a positive response bias, while /ɑ/ shows a reciprocal negative bias.

These results demonstrate that familiarization with a talker's voice *can* yield genuine improvement in the identification of individual vowels. But what *kind* of information is available in a series of syllables? One common hypothesis is that tokens of several vowels are needed to specify accurately the shape of a talker's vowel space (Ladefoged and Broadbent, 1957). In fact, some authors have specifically suggested that the "point vowels" /i, ɑ, u/ may be the primary calibrators of that space (Joos, 1948; Gerstman, 1968; Lieberman, Crelin, and

FIGURE 1

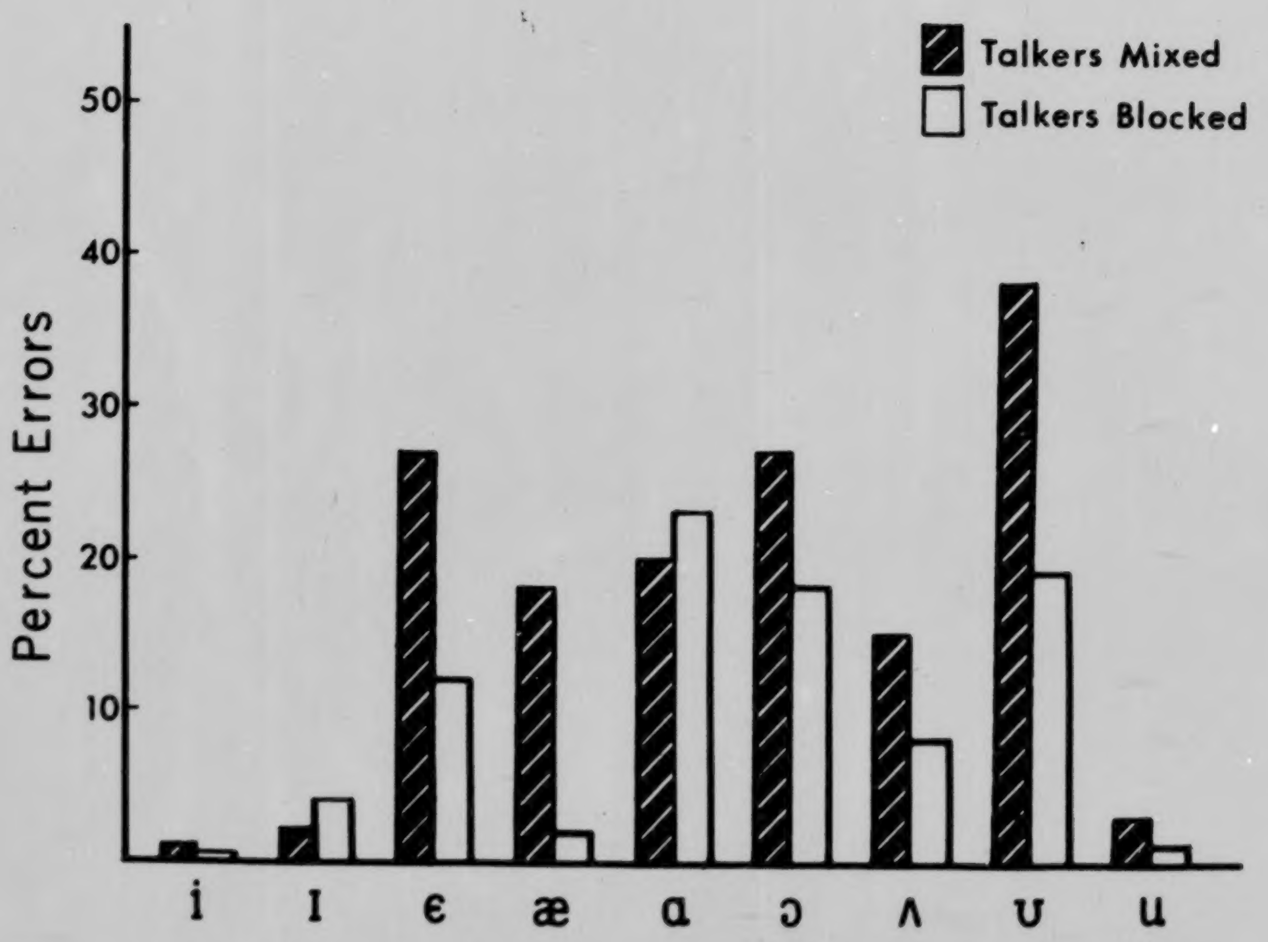


Figure 1: Mean percent errors in identification of each of nine vowels in /p–p/ environment.

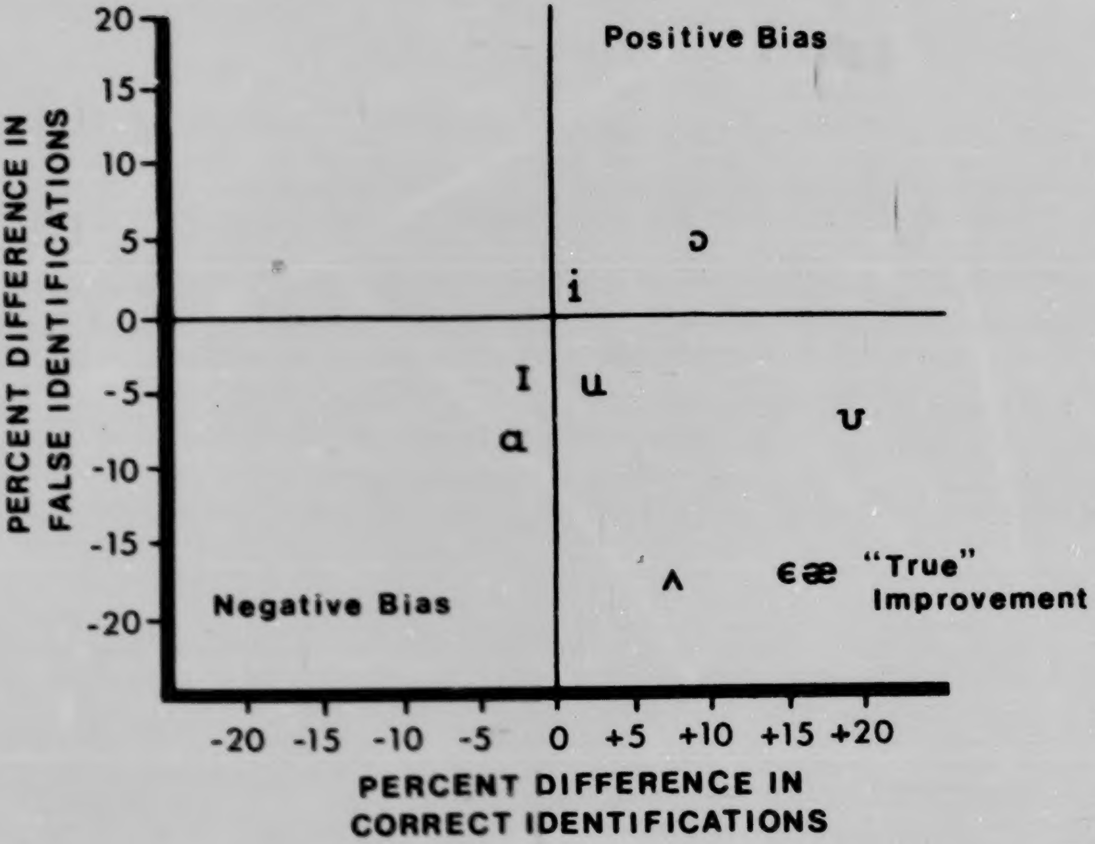


Figure 2: Changes in correct identifications and biases in vowel perception attributable to keeping the talker constant throughout a test.

Klatt, 1972). The point vowels occupy the corners of the vowel triangle and are said to have the most determinate vocal tract area functions (Stevens, 1973).

If the point vowels do serve as calibration signals for normalization, two predictions follow directly. First, experience with a talker's point vowels should substantially reduce errors in identifying ambiguous vowels. Second, experience with another set of three vowels should not be as effective in reducing errors, or not be effective at all.

To make a direct test of these predictions we adapted our earlier Mixed Talker Condition by preceding each test syllable with a set of three vowels spoken by the same person. We used two sets of vowel precursors: /hi, hɑ, hu/ and /hɪ, hæ, hʌ/. The vowels were spoken in /h-/ syllables to facilitate articulation, while minimizing nonvocalic sources of information. If point vowels *are* the source of familiarization effects, the data in the /hi, hɑ, hu/ precursor condition should resemble that in the Blocked Condition of the earlier study.

Figure 3 displays the overall percent errors in the two precursor conditions, along with our earlier results: the 9.3 percent error rate in the Blocked Condition, the 17 percent error rate in the Mixed Condition without precursors, and the results for the /hi, hɑ, hu/ precursors and the /hɪ, hæ, hʌ/ precursors. The point vowel precursors improved identification only slightly, reducing errors from 17.0 to 15.2 percent; the difference is *not* statistically significant by a *t* test. The three nonpoint vowels *also* reduced errors slightly, to 14.9 percent, though again the difference is not significant.

In other words, not only is there no evidence for a gain attributable to point vowels, but there is no difference between the point vowels and a set of nonpoint vowels. Overall, experience with specific sets of vowels seems to make little contribution to the total reduction of errors attributable to prior experience with a person's voice.

Before accepting these conclusions, it is worth checking whether there are improvements on specific vowels which are lost in the overall percentages. In Figure 4, the black columns indicate percent errors for each vowel in the Mixed Condition without precursors, the white columns are the results following the /hi, hɑ, hu/ precursors, and the hatched columns show errors following /hɪ, hæ, hʌ/. The point vowel precursors appear to help in identifying four vowels: /ɛ, ɔ, ʌ, ʊ/. Errors increase slightly for /æ/, and they increase on each of the precursor vowels, /i, ɑ, u/. This may be a kind of contrast effect between tokens of the vowel in the precursor string and in the test syllable itself, i.e., subjects may be biased away from choosing a point vowel. If so, it suggests that the pattern of change observed with point vowel precursors may reflect merely a shift in response biases and not a real change in identifiability.

Figure 5 plots change in correct responses against change in false identification; each axis represents the difference between the Mixed Condition *with* point vowel precursors and the Mixed Condition *without* precursors. The four vowels /ɛ, ɔ, ʌ, ʊ/ that showed apparent improvement in Figure 4 all appear in the upper right-hand quadrant of Figure 5--i.e., all four reflect a positive bias and none shows true improvement. Two other vowels, /ɑ, æ/, show a strong negative bias. A contrast effect with the precursor vowels is also evident, since each shows a negative response bias.

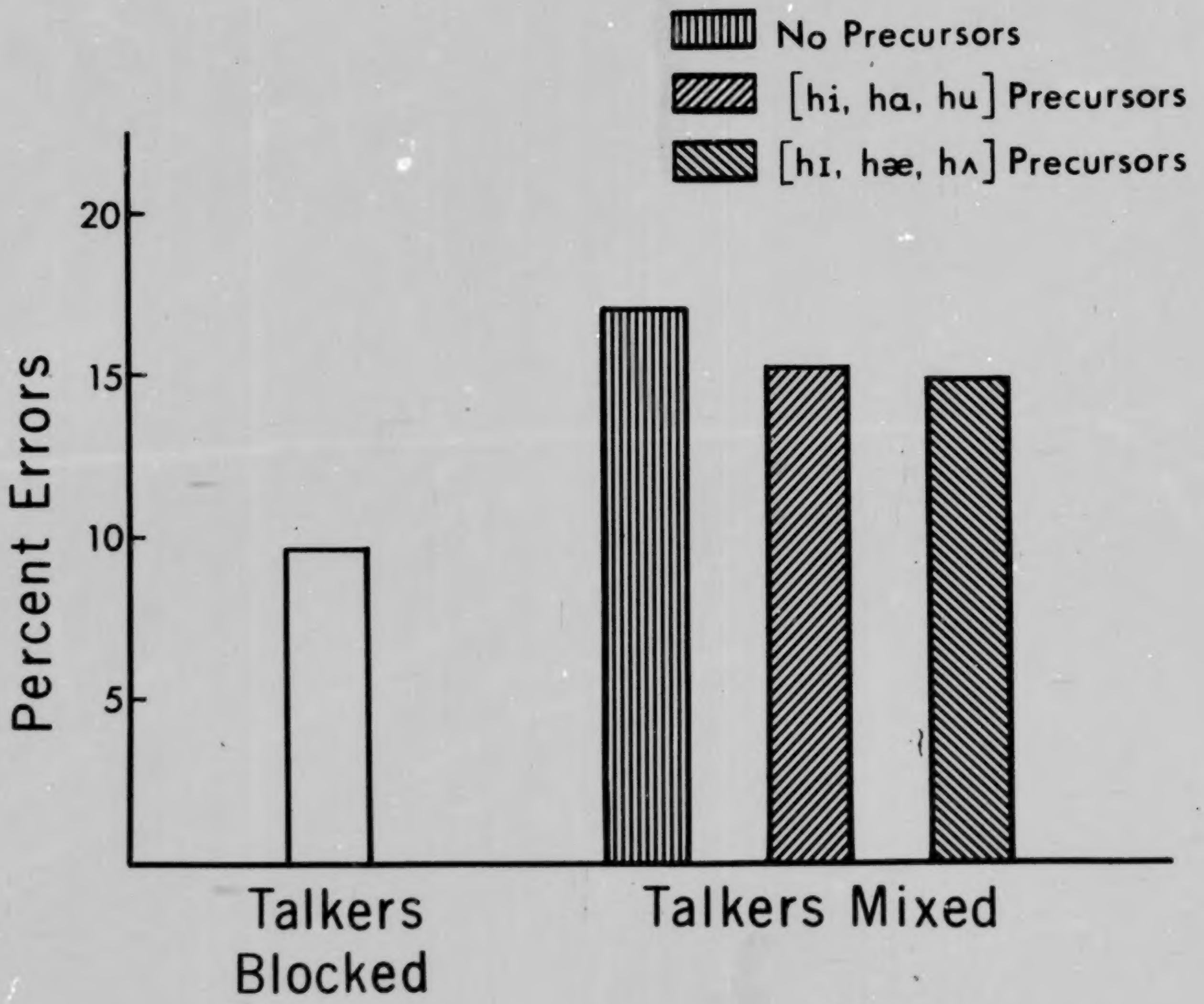


FIGURE 3

Figure 3: Mean percent errors in identification of nine vowels (averaged) in /p-p/ environment with and without precursors.

FIGURE 4

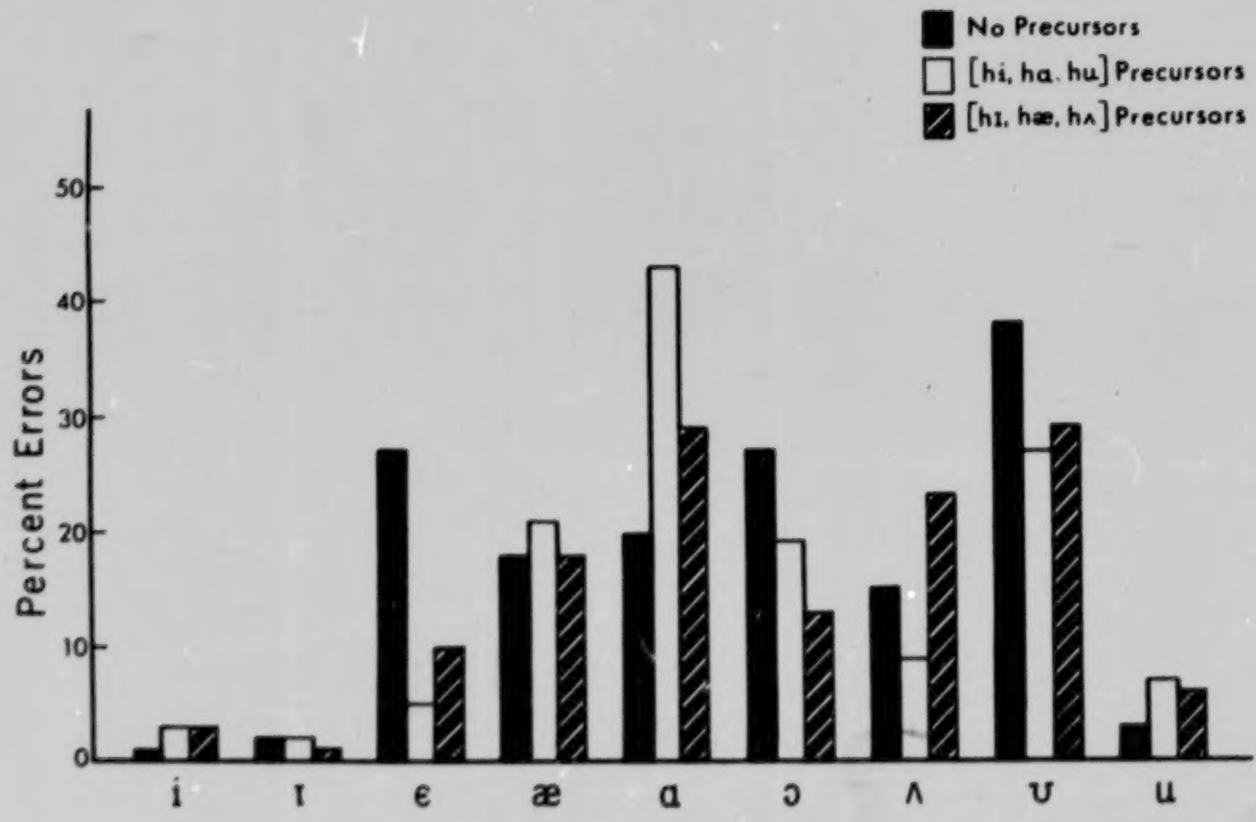


Figure 4: Mean percent errors in identification of each of nine vowels in /p-p/ environment with and without precursors.

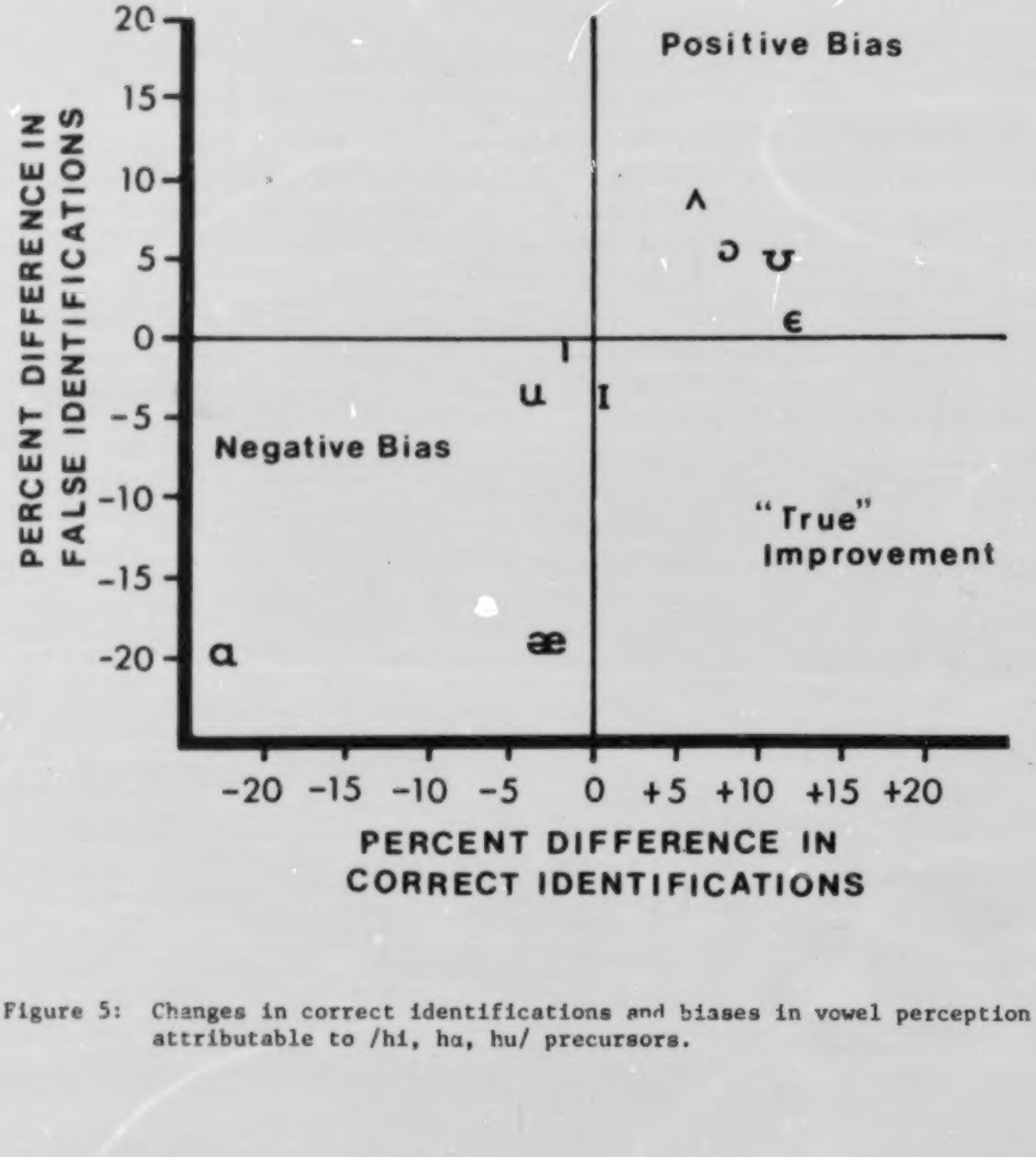


Figure 5: Changes in correct identifications and biases in vowel perception attributable to /hi, hɑ, hu/ precursors.

It is clear that experience with a talker's point vowels does have an effect on subsequent identifications. But in no case do our data demonstrate genuine improvement in perceiving ambiguous vowels. The effect seems limited to a biasing of response probabilities.

Overall, these data challenge the notion that extended experience with a talker's voice is the primary source of information about his vowels; and in particular, they challenge the notion that the point vowels play a special role as calibrators of a presumed vowel space.

REFERENCES

Gerstman, L. J. (1968) Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. *16*, 78-80.

Joos, M. (1948) Acoustic phonetics. Language *24* (Suppl.).

Ladefoged, P. (1967) The nature of vowel quality. In *Three Areas of Experimental Phonetics*. (London: Oxford University Press).

Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Amer. *29*, 98-104.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. *74*, 287-307.

Peterson, G. E. (1961) Parameters of vowel quality. J. Speech Hearing Res. *4*, 10-29.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of the vowels. J. Acoust. Soc. Amer. *24*, 175-184.

Stevens, K. N. (1973) The quantal nature of speech: Evidence from articulatory-acoustic data. In *Human Communication: A Unified View*, ed. by E. E. David and P. B. Denes. (New York: McGraw-Hill).

Stevens, K. N. and A. S. House. (1963) Perturbation of vowel articulations by consonantal context: An acoustical study. J. Speech Hearing Res. *6*, 111-128.

Consonant Environment Specifies Vowel Identity*

Winifred Strange,+ Robert Verbrugge,+ and Donald Shankweiler++

In the preceding paper, Verbrugge, Strange, and Shankweiler (1974) reported data on perception of nine English vowels spoken in a /p-vowel-p/ environment (the syllables, spoken by a panel of talkers, were recorded and assembled into a set of listening tests by randomly mixing the voices from token to token). A group of listeners, for whom these were novel voices, identified an average of 83 percent of the vowel nuclei as the intended vowels. This compares with an average of 91 percent correct identifications when the vowels were produced by the same talker. Thus, variation among talkers contributed less to vowel ambiguity than did other factors yet unidentified. Our intention here is to explore the effects on vowel perception of modifying the environment in which the vowels occur, when the talker is constant throughout a test and when the talker varies from token to token.

The influence of the nuclear vowel on the acoustic structure of the CVC (consonant-vowel-consonant) syllable is not confined to the steady-state middle portion, but may be traced throughout the whole temporal course of the syllable. We might therefore expect that transitions into and out of the steady-state target contain information that aids in specifying the vowel (Studdert-Kennedy, 1974). This possibility is strongly suggested by the little perceptual data available on the perception of isolated steady-state vowels. Fairbanks and Grubb (1961) found a strikingly high rate of misidentifications of isolated vowels which had been produced by phonetically-trained male talkers. Fujimura and Ochiai (1963), who compared identification of Japanese vowels spoken in syllabic

---

*Paper presented at the 87th meeting of the Acoustical Society of America, New York, 25 April 1974.

+University of Minnesota, Minneapolis.

++Haskins Laboratories, New Haven, Conn., and University of Connecticut, Storrs.

Acknowledgment: This paper and the preceding one report a portion of research begun during the academic year 1972-73 while D. Shankweiler was a guest investigator at the Center for Research in Human Learning, University of Minnesota, Minneapolis. The work was supported in part by a grant from the National Institute of Child Health and Human Development to the Center, and in part by grants awarded to Shankweiler and to J. J. Jenkins by the National Institute of Mental Health. We wish to thank Thomas Edman and Kevin Jones for their assistance in every phase of the experimental work and James Jenkins for his advice and encouragement.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

context with identification of segments gated out of the vowel centers, showed that identifications shifted in the absence of flanking transitions.

Thus, we recognize two sources of variation in vowels that may affect their identifiability. First, there is the variation due to talker differences. We can assess the importance of this source of variation by using the Mixed Talker and Blocked Talker Conditions described in the preceding paper (Verbrugge et al., 1974). Second, there is variation associated with consonantal environment (or its absence). We can assess the importance of this source by constructing tests with vowels in a consonantal frame and isolated steady-state vowels. A direct comparison of the effects on vowel perception of these two factors can be made by varying both independently in an experiment in which medial vowels and isolated vowels are each presented for identification in both a Mixed Talker test and a Blocked Talker test.

Vowels spoken in isolation were obtained from the same panel of talkers who produced vowels in the /p-vowel-p/ environment for the previous experiments.[1] The tests were constructed in the same way as before: 15 talkers produced isolated vowels for the Mixed Talker Condition and the same man, woman, and child each produced the full series of vowel tokens for the Blocked Talker Condition.

Figure 1 shows the results of the isolated vowel tests, along with the earlier results for the vowels in the /p-p/ environment. Overall, there were many more errors in identifying isolated vowels than there were in identifying medial vowels. This is true both when the talker varied from token to token (Mixed Condition) and when the talker was constant throughout the test (Blocked Condition). Incorrect identifications of isolated vowels produced by the entire panel (Mixed) averaged 42 percent as compared to 17 percent for medial vowels. Errors in identification of tokens produced by the three prototypic talkers (Blocked) averaged 31 percent on isolated vowels and only 9 percent on medial vowels.

We may now consider why vowels in a consonantal environment are identified so much more accurately than isolated vowels. One possibility is that transitions play a role in determining a talker's vowel space. Since the loci of formant transitions for a particular consonant are a function of vocal tract size and shape, transitions might serve as additional calibration signals for vowel normalization. The informative value of these transitions could explain the difference between errors in identification of isolated vowels and of vowels in consonant environment for tests in which the talker was different from token to token, as in the Mixed Condition. However, when a single talker produces all the tokens on a test, there is no necessity for repeated recalibration; therefore, the presence of transitions might be expected to have little effect on vowel identification. That is, when talkers are blocked we should expect a similar error rate for steady-state vowels with and without consonantal environment. But it may be seen from Figure 1 that medial vowels are much more accurately identified than isolated vowels even when talker variation is not a factor. Indeed, the

---

[1] Vowels were indicated to the talkers by printed cards which specified "vowel as in peep," "vowel as in pip," etc. The talkers were not provided with spoken models.

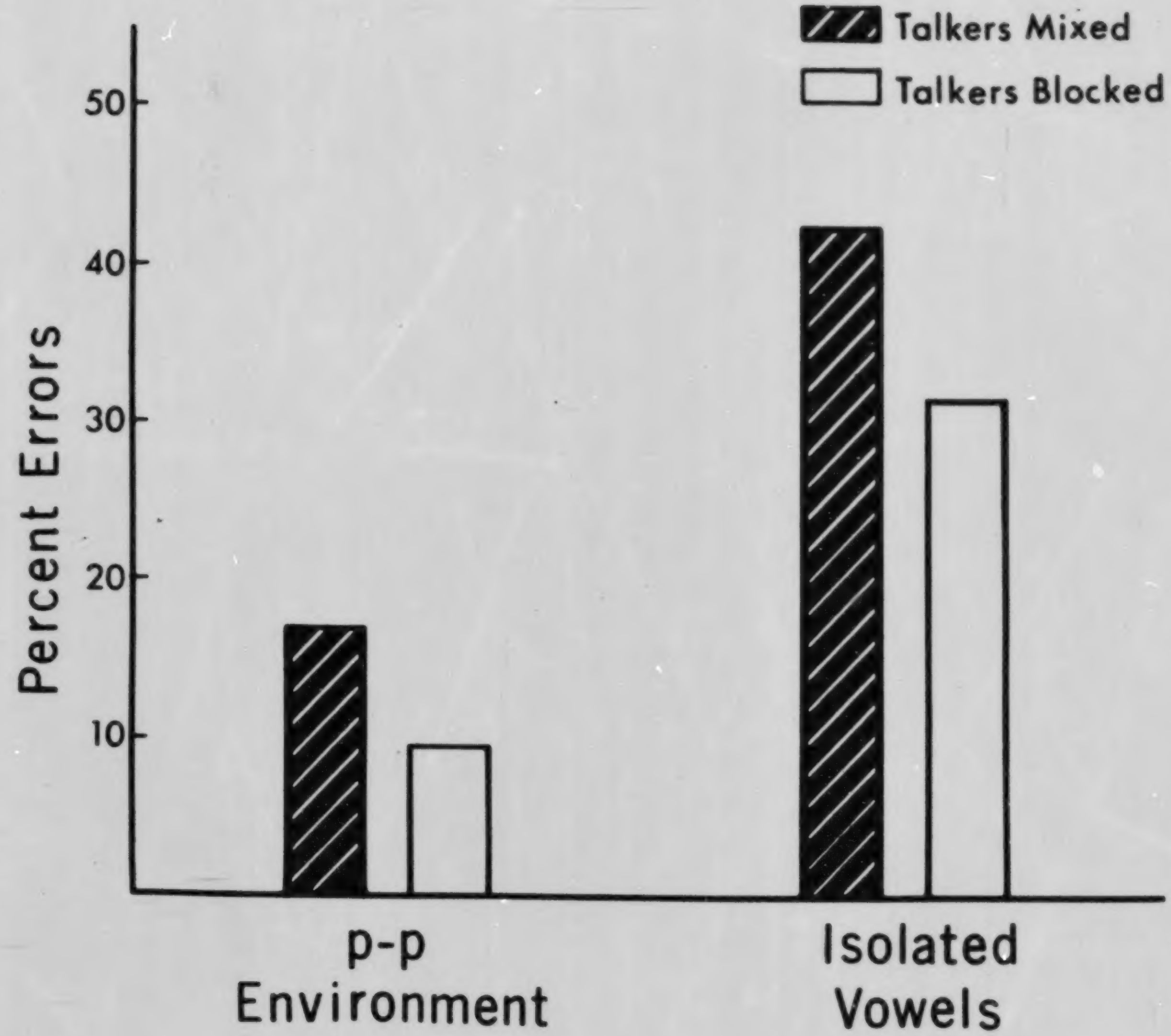


FIGURE 1

Figure 1: Mean percent errors in identification of nine vowels (averaged) in /p-p/ environment and in isolation.

effect of talker variation is roughly constant whether a consonantal environment is present or not. These results indicate that the effect of environment is more important than the effect of talker variation. Absence of consonantal transitions accounts for a much larger proportion of misidentified vowels than does talker differences.

A vowel-by-vowel analysis of errors indicates that the presence of consonantal transitions produced a consistent facilitation in identification of all nine vowels. Figure 2 shows the data for tests on which talkers were blocked. It is noteworthy that the advantage of the consonantal environment is clearly seen for every vowel. The same is true of the results of tests on which the talkers were mixed, as shown in Figure 3. Note that the superior identifiability of /i/ and /u/, the extreme points in vowel space, is not convincingly retained in the absence of transitions. This leads us to question the hypothesis that maintains that these vowels are good perceptual targets because they uniquely specify vocal tract configuration (Lieberman, Crelin, and Klatt, 1972).

The idea that consonant transitions make their contribution to perception of the vowel by providing cues for normalization is not supported. We may suppose that the efficacy of the /p-vowel-p/ environment in aiding identification of the enclosed vowels has nothing to do with normalization. We must seek another explanation of the advantage to vowel perception conferred by the consonantal environment.

It is clear from our data and those of Fairbanks and Grubb (1961) that isolated, steady-state vowels, although they presumably conform closely to the idealized target vowels of phonetic theory, are extremely poorly specified targets from the standpoint of the perceiver. Lehiste and Peterson (1959) have shown that many hours of familiarization with a particular talker's vowels are required to yield high accuracy in identification of vowels in isolation. Thus it may be that categorization of these vowels is a rather special ability which may have little bearing on the processes of speech perception in a natural setting.

A final set of experiments evaluates the generality of the notion that the consonantal environment provides critical information for the identification of medial vowels. If provision of an environment aids in perception of the vowel only when the consonantal frame is fixed, the finding would be of limited interest. If listeners, by knowing the phonemic identity of the consonant beforehand, somehow "work backward" from that knowledge to decode the steady-state portion of the syllable, we might expect that the identification of vowels in a variable consonantal frame would be less accurate than for vowels in a fixed environment. If, on the other hand, the acoustic specification of vowels, like consonants, is contained in the dynamic configuration of the syllable, we might expect listeners to utilize these dynamic cues whether or not the identity of the consonants is known in advance.

We generated a set of syllables in which the nine vowels were produced in varying consonantal environments. The six stop consonants were paired with the nine vowels such that each consonant preceded and followed each vowel an equal number of times. These syllables were spoken by a panel of talkers and presented to listeners with talkers randomized (Mixed Talker Condition). Listeners were asked to identify only the vowel in each syllable.

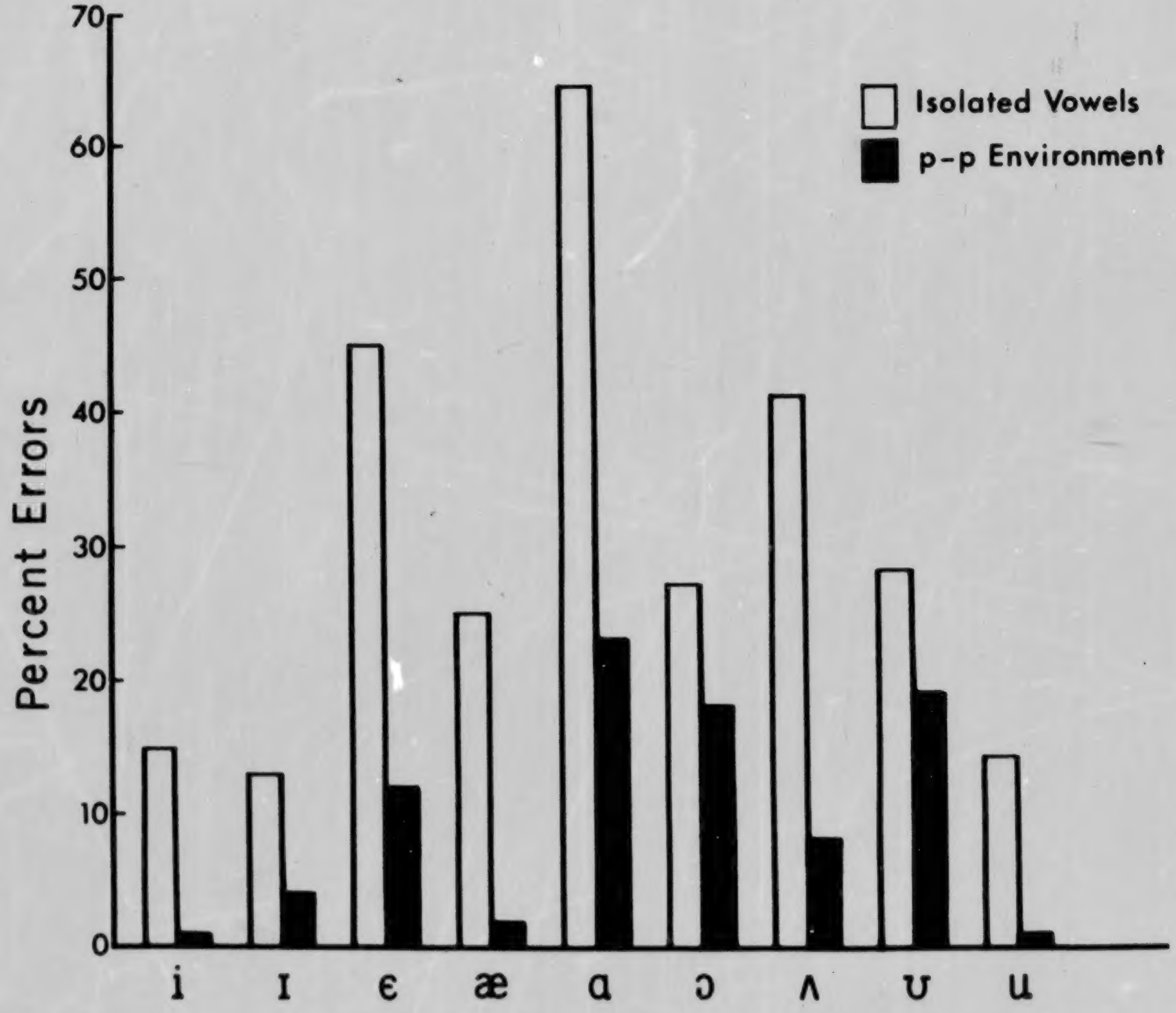


Figure 2: Mean percent errors in identification of each of nine vowels in isolation and in /p-p/ environment: talkers blocked.

FIGURE 2

FIGURE 3

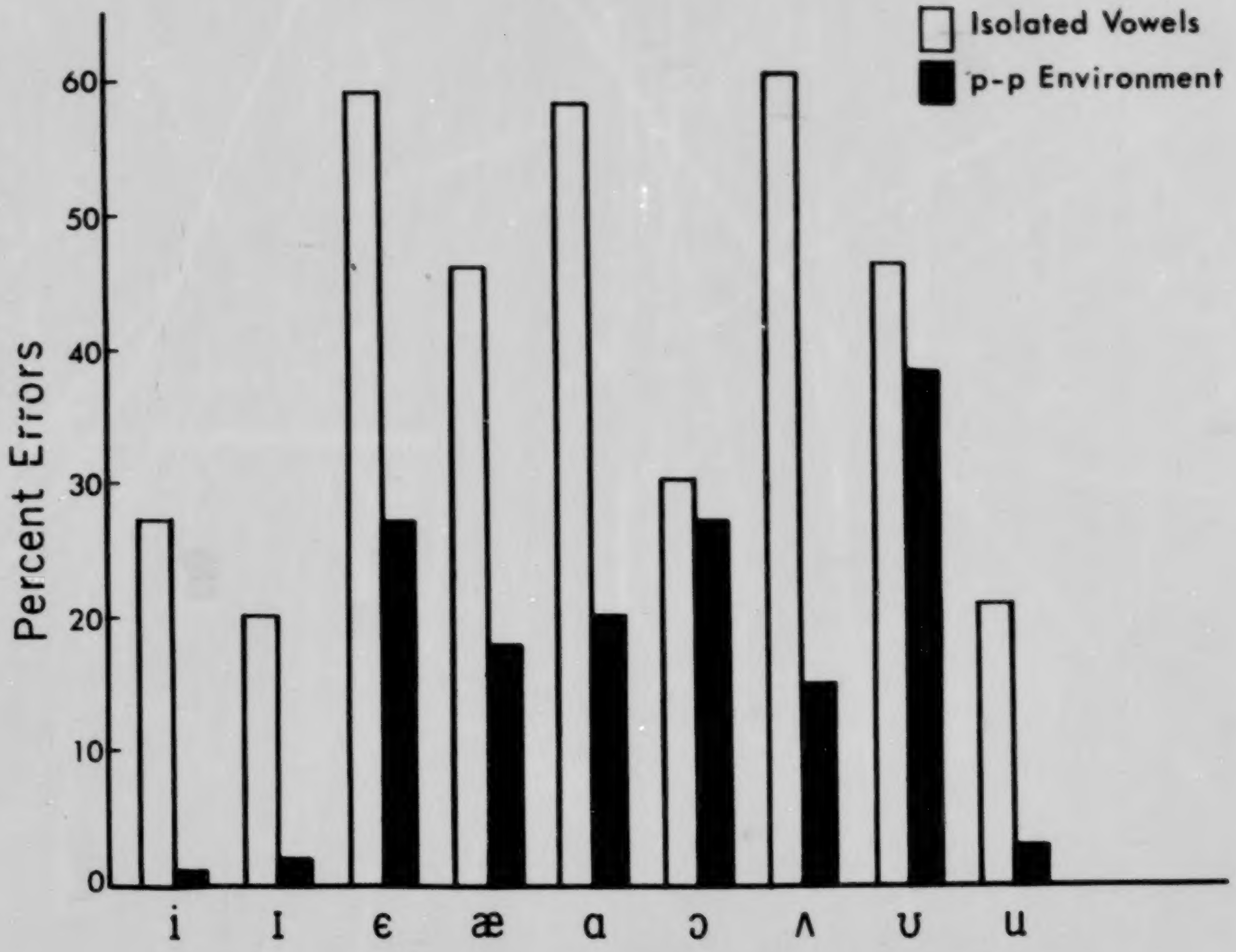


Figure 3: Mean percent errors in identification of each of nine vowels in isolation and in /p-p/ environment: talkers mixed.

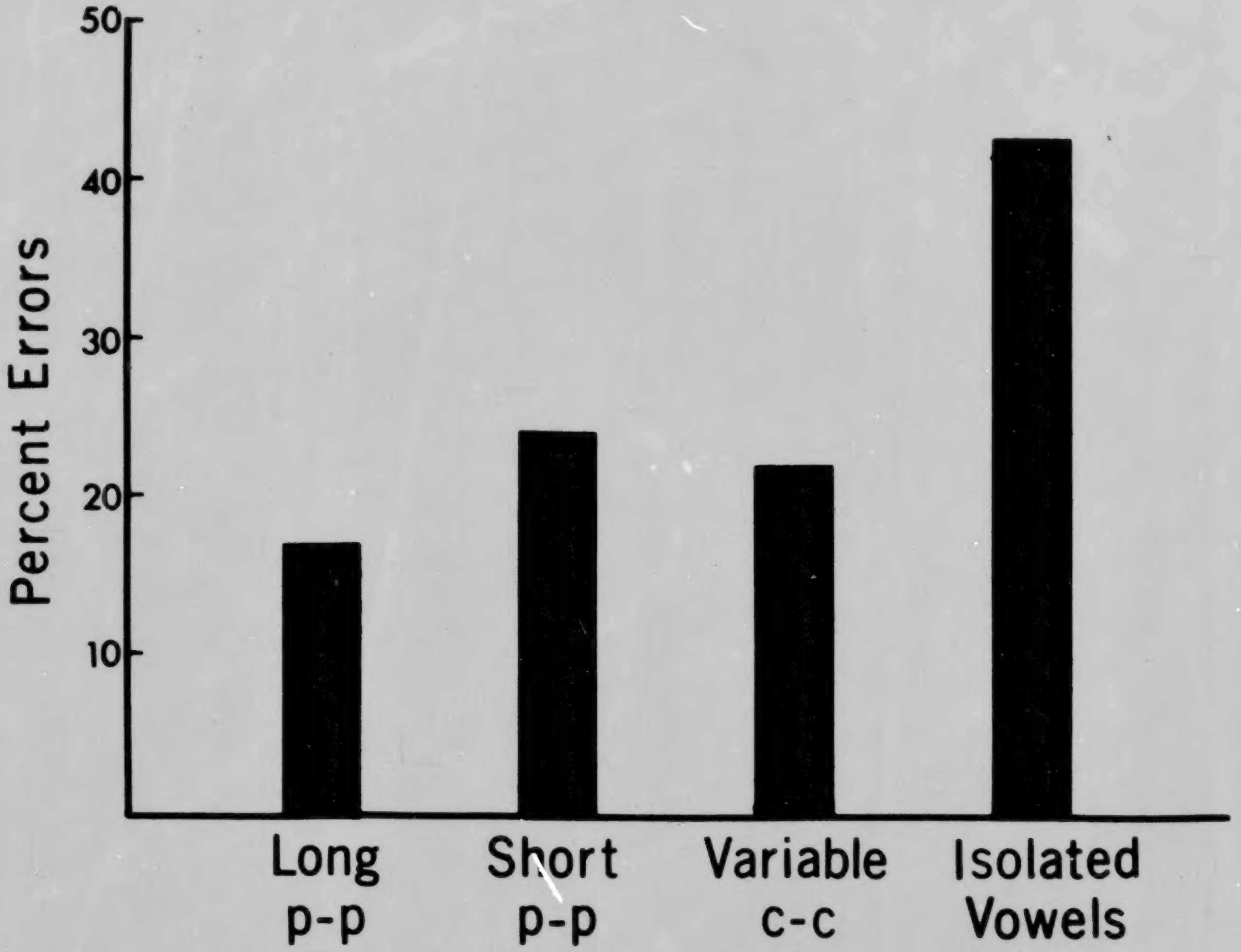


FIGURE 4

Figure 4: Mean percent errors in identification of nine vowels (averaged) in four environments: /p-p/ syllables produced in citation form, /p-p/ syllables excised from sentences, /c-c/ syllables in which consonantal frame varies, vowels spoken in isolation.

Figure 4 presents the results of identification tests conducted with vowels in a variable consonantal frame (labeled "variable c-c"), in a /p-p/ frame (labeled "long p-p"), and in isolation. It may be seen that although the variable frame gives rise to slightly more errors than the fixed frame (21 percent and 17 percent, respectively) the identification of vowels in isolation is considerably less accurate than for vowels in randomly varying environments, 42 percent as compared to 21 percent.

In a final experiment, we tested the identifiability of vowels for syllables in which the formants failed to reach their steady-state targets. To measure the identification of these "reduced" vowels, we had our panel of talkers produce the /p-vowel-p/ syllables in a fixed sentence frame, with the target syllables in nonstressed sentence position. The syllables were then excised from the sentences and presented to listeners in the same manner as were the /p-vowel-p/ syllables that had been produced in citation form.

The results are summarized in the second bar of Figure 4 (labeled "short p-p"). Accuracy of identification even of rapidly articulated vowels exceeds by a considerable margin the rate achieved for isolated vowels (23 percent compared to 42 percent).

In conclusion, we have shown that providing a consonantal frame increases the likelihood of correct identification of medial vowels. This is true not only when each vowel is placed in a fixed environment, but also when changing consonantal environments and rapid rates of articulation greatly increase the complexity and variability of the acoustic information that specifies a vowel. These variations, which add greatly to the complexity of any physical description of the signal, produce little effect upon perceptual judgments. Indeed, the complexities introduced by syllabic structure better serve the requirements of the perceptual apparatus than do simple steady-state targets. These data lead us to emphasize that much remains to be learned about what specifies a vowel. The solution is to be sought in the dynamic syllable, not in an idealized steady-state target.

REFERENCES

Fairbanks, G. and P. Grubb. (1961) A psychophysical investigation of vowel formants. J. Speech Hearing Res. 4, 203-219.

Fujimura, O. and K. Ochiai. (1963) Vowel identification and phonetic contexts. J. Acoust. Soc. Amer. 35, 1889(A).

Lehiste, I. and G. Peterson. (1959) The identification of filtered vowels. Phonetica 4, 161-177.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. 74, 287-307.

Studdert-Kennedy, M. (1974) The perception of speech. In Current Trends in Linguistics, ed. by. T. A. Sebeok. (The Hague: Mouton). [Also in Haskins Laboratories Status Report on Speech Research SR-23, 15-48 (1970).]

Verbrugge, R., W. Strange, and D. Shankweiler. (1974) What information enables a listener to map a talker's vowel space? Haskins Laboratories Status Report on Speech Research SR-37/38 (this issue).

Identification of Vowel Order: Concatenated Versus Formant-Connected Sequences

M. F. Dorman,[+] James E. Cutting,[++] and Lawrence J. Raphael[+]
Haskins Laboratories, New Haven, Conn.

In order for listeners to identify accurately the temporal order of a sequence of four concatenated vowels, each vowel must be between 125 and 200 msec in duration (Warren, 1968; Thomas, Hill, Carrol, and Garcia, 1970). These data fit nicely into a theory of speech perception (Massaro, 1972) that suggests that the perceptual processing of vowels lasts between 120 and 250 msec.

The perception of concatenated speech, however, bears only a slight resemblance to the perception of speech in connected discourse. For example, in discourse listeners can follow speech at rates up to 400 words per min, or approximately 30 phonemes per sec (Orr, Friedman, and Williams, 1965). In marked contrast to the results of Warren (1968) and Thomas et al. (1970), these data suggest that speech signals can be correctly ordered when the average phoneme duration is only 30 to 40 msec. The discrepancy between these two estimates suggested to us that concatenated sequences produce a spuriously high estimate of vowel duration necessary for the temporal ordering of speech. This observation prompted a series of experiments comparing temporal order judgments for concatenated and coarticulated sequences of vowels.

EXPERIMENT I

Method

Warren-type repeating sequences were generated on the Haskins Laboratories' parallel resonance synthesizer. The sequences consisted of long steady-state vowels (V_1) [i, æ, ɔ, u] and of consonant-vowel-consonant (CVC) syllables [bib, bæb, bɔb, bub]. All stimuli were 120 msec in duration, well below the critical duration (168 msec) noted by Thomas et al. (1970) for 75 percent correct performance on synthetic vowels. All stimuli had the same fundamental frequency (110 Hz) and overall amplitude contour. Initial and final formant transitions in the CVCs were 45 msec in duration, leaving 30 msec of steady-state vowel. Stimuli within the same class (V_1 and CVC) were permuted in the six possible orders. These sequences were recorded on audio tape with 10 msec between successive items. Each sequence began at a very low volume, gradually increased in volume over the course of 5 sec to a maximum intensity (approximately 80 db), remained at that

---

[+] Also Herbert H. Lehman College of the City University of New York.

[++] Also Yale University, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

maximum for 10 sec, and then decreased to its original low volume during the final 5 sec period. Stimulus class and sequence orders were randomized for presentation to the listeners.

Twenty-two Yale undergraduate students participated in the task as part of a course requirement. The stimuli were reproduced on an Ampex AG500 tape recorder via an Ampex 620 loudspeaker. Tokens of the steady-state vowels at 2 sec durations were played to the listeners until they could accurately identify the vowels. The listeners were then told that they would hear more rapid vowel sequences and CVC sequences, and were instructed to report the identity of the vowels in the order that they were heard (disregarding the /b/s in the CVC stimuli).

Results and Discussion

Table 1 shows the averaged performance of the listeners for V_1 and CVC stimuli for each of the six orders. Notice that in terms of performance summed over all sequence orders, there was no significant difference between the two classes of stimuli: V_1 and CVC stimulus orders were correctly identified on 67 and 70 percent of all trials, respectively. This result is deceptive, however, since it sums over sequence orders with quite varied performance levels. Two orders, number 2 [i, æ , u, ɔ] and number 4 [i, ɔ, u, æ] were more difficult to identify than the other four. The results of these two orders yielded the most interesting pattern: V_1 sequences were identified on only 51 percent of such trials, and CVC sequences were identified on 64 percent. This pattern occurred against a background of small differences between the two stimulus classes on the other four sequence orders: 75 percent correct for V_1 stimuli, and 73 percent for CVCs. Listeners readily volunteered that sequence orders number 2 [i, æ , u, ɔ] and number 4 [i, ɔ, u, æ] were more difficult to identify than the others, describing the difficulty in terms of the sequences "flying apart." We recognized this as the hallmark of "auditory streaming."

TABLE 1

		V_1	CVC	Average
1)	i æ ɔ u	86	80	83
2)	i æ u ɔ	45	57	51
3)	i ɔ æ u	84	69	76
4)	i ɔ u æ	57	70	63
5)	i u æ ɔ	60	67	64
6)	i u ɔ æ	70	77	74
	Average	67	70	

Bregman and Campbell (1971) have reported that when listeners are presented a repeating sequence of six brief (100 msec) tones which alternate between high and low frequencies, listeners are unable to report correctly the high-low sequence, reporting instead two streams of tones, one containing the high tones and the other containing the low tones. Within a stream, ordering is reasonably accurate (73 to 79 percent), however between streams ordering is no better than chance. Bregman and Campbell have termed this phenomenon "primary auditory stream segregation." The perceptual experience in the present study of the vowel sequences "flying apart" on sequences 2 and 4 appears identical to the auditory stream segregation of Bregman and Campbell's nonspeech auditory signals. Although sequences 2 and 4 were characterized by physically separated [i]-[u] and [æ]-[ɔ] pairs, listeners reported hearing each as a unit pair. This phenomenon may be accounted for in terms of perceptual streaming of the first formants, the most prominent and lowest-frequency component of the four vowels. The first formant (F1) frequency value for both [i] and [u] was 286 Hz, whereas for [æ] and [ɔ] it was 666 Hz and 614 Hz, respectively. Since [i] and [u] have the lowest-frequency first formants, these vowels appear to be heard as one stream, and [æ] and [ɔ] with higher first formants appear to form a separate stream, as suggested in the top panel of Figure 1. Only sequence orders 2 and 4 meet the requirement of having alternating high and low first formants, and indeed these are the orders that were most difficult for listeners to identify.

A second important observation is that the sequence order of 30 msec vowels in the context of initial and final [b] could be identified at least as accurately as the 120 msec vowels. Since 30 msec is far below the vowel duration necessary for accurate temporal ordering of concatenated vowels (Thomas et al., 1970), the formant transitions in the CVC sequences may act in a manner similar to silence in facilitating recognition of vowel sequence order. Warren (1968), for example, has reported that, although sequence orders of four concatenated 200-msec vowels are very difficult to perceive, the same orders with 150 msec vowels separated by 50 msec of silence are relatively easy to perceive.

Since the results of the present study suggest that certain sequence orders are more difficult to identify than others, Experiment II was designed, in part, to observe such differences in greater detail. In addition, Experiment II was designed to compare the relative contribution of transitions and of silence to the perception of temporal order.

EXPERIMENT II

Method

The V_l and CVC stimuli from the previous experiments were used again. In addition, short, steady-state vowel stimuli (V_s) were synthesized. They were identical to the V_l stimuli in all respects except duration. Whereas the long vowels were 120 msec in duration, the short vowels were only 30 msec long, and thus identical to the steady-state vowel portion of the CVC stimuli. The other 90 msec of the stimuli was replaced by silence. Again, all stimuli within a class were permuted in the six possible sequence orders, but the two most difficult orders (numbers 2 and 4) were represented twice as often as the other four. Schematic spectrograms of the V_l, V_s, and CVC stimuli in the order [i, ɔ, u, æ] are shown in Figure 1. Each sequence was recorded in the same fashion as in Experiment I, and class of stimuli and sequence order were randomly intermixed.

Eight students at Herbert Lehman College of the City University of New York and three staff members from Haskins Laboratories served as listeners. Stimuli were reproduced for the Lehman College listeners on a Revox 1122 tape recorder via an AR-4x loudspeaker, and for the Haskins Laboratories listeners on the same apparatus as in Experiment I.

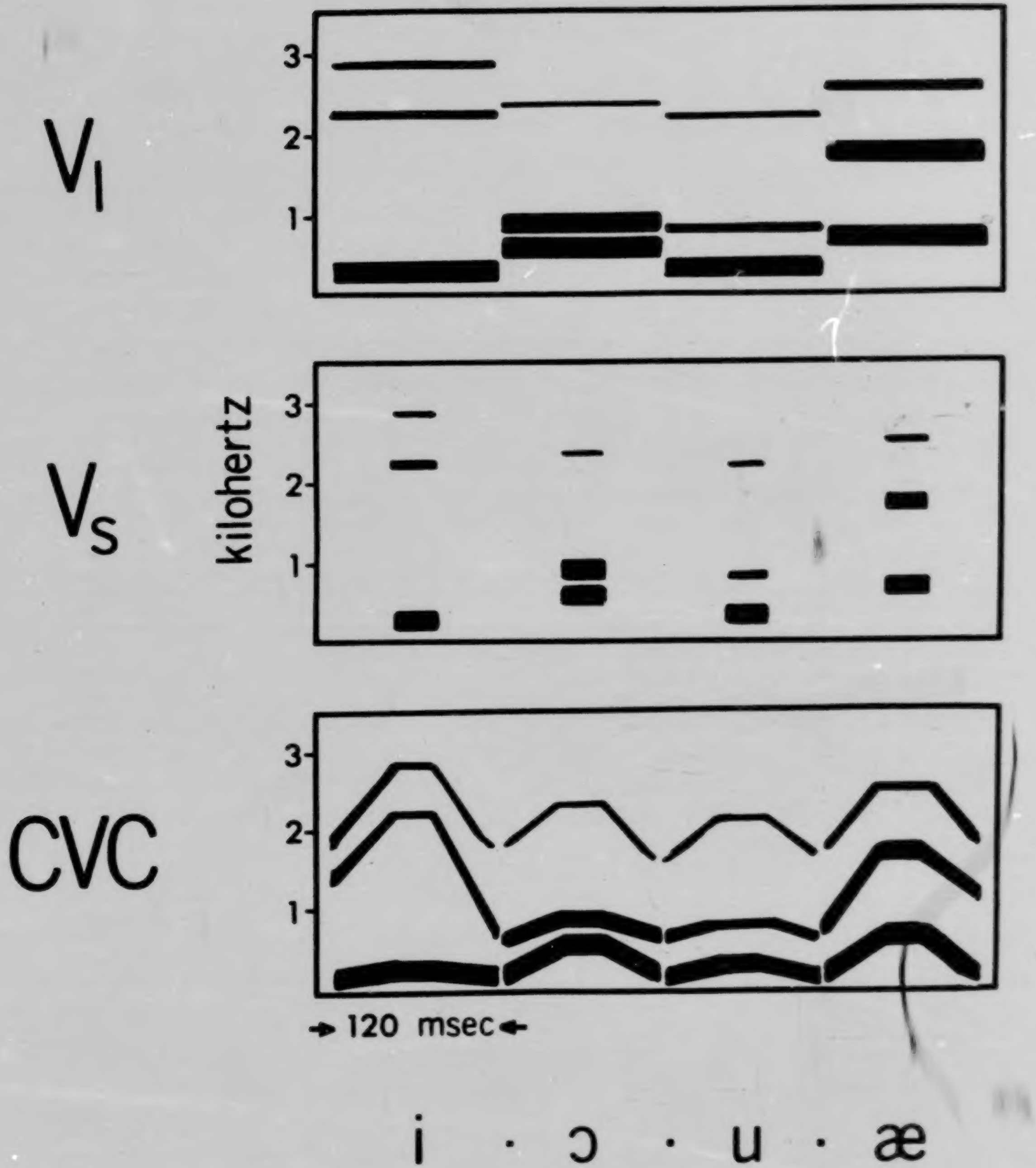


Figure 1: Schematic spectrograms of long-vowel (V_l), short-vowel (V_s), and consonant-vowel-consonant (CVC) sequences.

Results and Discussion

As shown in Table 2, V_1 and V_s sequences were considerably more difficult to order than the CVC sequences, with their respective average performance levels at 52, 47, and 71 percent. CVC sequence orders were significantly easier to identify than those of either the V_1 stimuli ($T(8) = 3$, $p < .05$) or the V_s stimuli ($T(8) = 5$, $p < .05$). The differences are even more striking for the two difficult sequence orders, numbers 2 and 4: such V_1 sequences were correctly identified on only 42 percent of all presentations, V_s sequences were identified on 44 percent, and CVC stimuli on 73 percent. For the other sequences (1, 3, 5, and 6) temporal-order accuracy was 63, 50, and 69 percent for the V_1, V_s, and CVC sequences, respectively. There were no significant differences among the sequences of V_1 and V_s stimuli. There were also no systematic differences between the Haskins and Lehman subjects.

TABLE 2

		V_1	V_s	CVC	Average
1)	i æ ɔ u	88	50	88	75
*2)	i æ u ɔ	44	38	78	53
3)	i ɔ æ u	50	13	50	38
*4)	i ɔ u æ	39	50	67	52
5)	i u æ ɔ	63	75	88	75
6)	i u ɔ æ	50	63	50	54
	Average	52	47	71	

*represented twice as often as other orders.

The perceptual advantage of the CVC sequences over the two classes of vowels stems primarily from performances on the sequence numbers 2 and 4. This outcome suggests that the reason there was no significant difference between the CVC and V_i stimuli in Experiment I was a ceiling effect induced by the over-representation of the easier sequence numbers 1, 3, 5, and 6. The perceptual advantage of the CVC sequences over the V_s sequences suggests that tracking formants between vowels is more effective than silence in reducing auditory streaming. Such an outcome is encouraging since few silent intervals occur in running speech, yet correct phoneme order is effortlessly extracted.

The superiority of the CVCs over both types of vowel sequences suggests further that there may be other ways of perceptually "gluing" the vowels together. Experiment III was designed, in part, to determine if streaming could be overcome through the use of long, continuous transitions between vowel nuclei. This tactic has proved useful in limiting primary auditory stream segregation in sequences of pure tones (Bregman and Dannenbring, 1973). Experiment III was also designed to determine if streaming is suppressed by all transitions, or only by transitions that are phonetically possible.

EXPERIMENT III

Method

The V_1 and CVC stimuli were used again. In addition, two other sets of stimuli were generated. Both sets contained 30-msec steady-state vowel segments corresponding to the vowels [i, æ , ɔ, u] and both had initial and final formant transitions. In one set the transitions were context-dependent, gliding gradually over the course of 90 msec from the steady-state formant values of one vowel into the succeeding vowel. Because of the connecting transitions, they are termed V_t stimuli. The second set, termed C'VC' stimuli, contained most of the features of CVCs except that the formant transitions were turned upside down; that is, instead of all transitions gliding upwards into the vowel and downwards after it (appropriate for the perception of [b]), all transitions glided downwards into the vowel and back upwards after it (inappropriate for the perception of any consonant phoneme). Only three of the six possible stimulus orders were selected: orders 2 and 4 to optimize error probability, and order number 1 for comparison purposes. V_1, V_t, CVC, and C'VC' sequences of [i, ɔ, u, æ] are shown in Figure 2. Class of stimuli and sequence order were randomized and recorded on audio tape.

The listeners were nine Lehman College undergraduate students. Stimuli were reproduced on a Revox 1122 tape recorder via an AR-4x loudspeaker. In all other respects the procedure was the same as in the two previous studies.

Results and Discussion

As shown in Table 3, there was a large difference between the accuracy of sequence identification for the two types of vowel stimuli. V_1 sequence orders were identified on only 30 percent of all presentations, whereas V_t orders were identified on 65 percent. All subjects demonstrated this difference ($\underline{T}(9) = 0$, $\underline{p} < .01$). The CVC sequences were again identified more accurately than the long vowels ($\underline{T}(9) = 0$, $\underline{p} < .01$), and again all listeners demonstrated this effect. There was no difference between CVC and V_t sequence-order identification. C'VC' sequence orders were essentially incomprehensible, and were identified at a chance performance level.

TABLE 3

	V_1	V_t	CVC	C'VC'
1) i æ ɔ u	50	55	44	11
2) i æ u ɔ	6	78	83	17
4) i ɔ u æ	33	61	55	11
Average	30	65	61	13

CONCLUSIONS

Two trends are prominent in the data of the present study. First, the employment of gradual transitions between the 30-msec vowel nuclei was successful

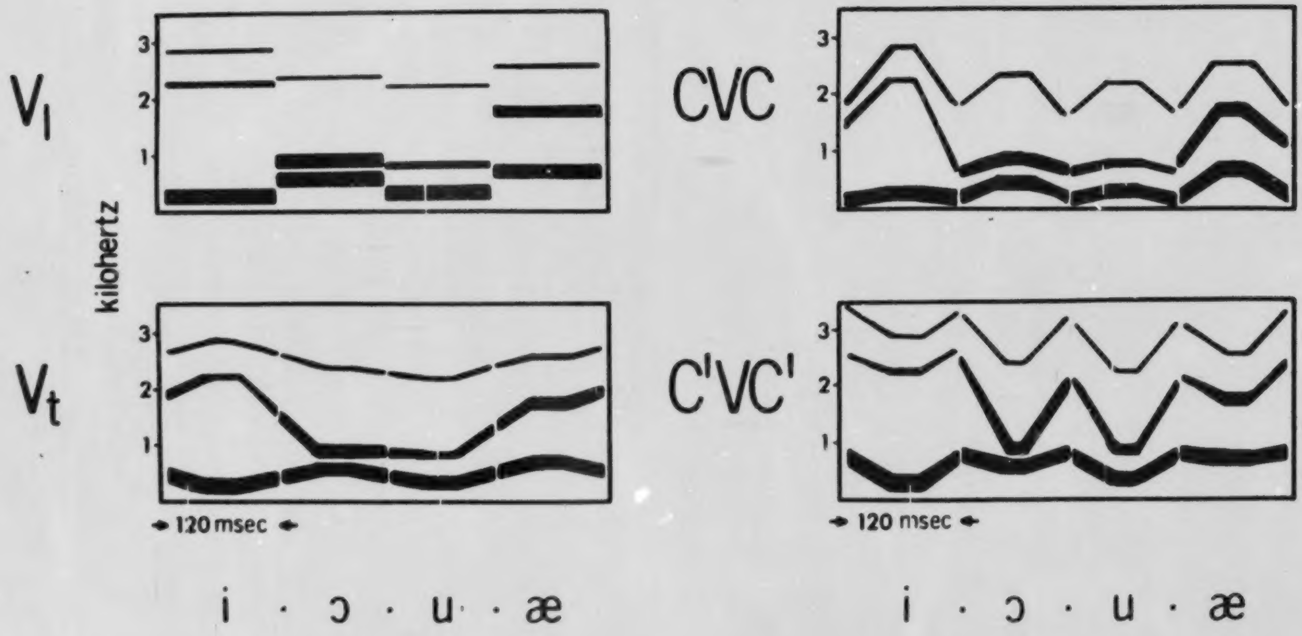

V$_l$
V$_t$
CVC
C'VC'
kilohertz
3
2
1
120 msec
i · ɔ · u · æ
i · ɔ · u · æ

FIGURE 2

Figure 2: Schematic spectrograms of long-vowel (V_l) and vowel-with-transition (V_t) sequences; consonant-vowel-consonant (CVC) sequences; and sequences with phonetically impossible transitions (C'VC').

in "gluing" back together the previously "streamed" vowel sequence orders. Such transitions are reasonable in an articulatory sense, just as the transitions in the CVC sequences are also reasonable. This points out the second trend: only those transitions that are phonetically possible were successful in limiting auditory streaming. Articulatorily impossible transitions such as those in the C'VC' stimuli failed to limit streaming. In fact, they decreased the probability of perceiving correct stimulus order.

From the results of Experiments I-III, we conclude (a) that certain sequence orders of four vowels are more difficult to perceive than others, (b) that the difficulty in perceiving these sequence orders is intimately related to the phenomenon of auditory stream segregation, (c) that streaming cannot be eliminated by replacing most of the vowel with silence, but that it can be virtually eliminated by replacing most of the vowel with formant transitions appropriate for the stop consonant [b] or with formant transitions that link the vowel nuclei with one another, and (d) that the suppression of auditory streaming is possible only through the use of transitions corresponding to gestures that could be articulated. In other words, the more the repeating sequences resemble connected discourse, the more facile listeners are at identifying temporal order. Phonemes in connected discourse do not stream precisely because they are coarticulated and not concatenated.

REFERENCES

Bregman, A. S. and J. Campbell. (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. J. Exp. Psychol. 89, 244-249.

Bregman, A. S. and G. L. Dannenbring. (1973) The effect of continuity on auditory stream segregation. Percept. Psychophys. 13, 308-312.

Massaro, D. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.

Orr, D. B., H. L. Friedman, and J. C. Williams. (1965) Trainability of listening comprehension of speeded discourse. J. Educ. Psychol. 56, 148-156.

Thomas, I. B., P. B. Hill, F. S. Carrol, and B. Garcia. (1970) Temporal order in the perception of vowels. J. Acoust. Soc. Amer. 4, 1010-1013.

Warren, R. M. (1968) Relation of verbal transformations to other perceptual phenomena. In IEE/NPL Conference on Pattern Recognition, Conference Publication No. 42, Suppl. (Teddington, England: Institution of Electrical Engineers).

On "Explaining" Vowel Duration Variation*

Leigh Lisker+
Haskins Laboratories, New Haven, Conn.

As they go about investigating physical aspects of speech communication, phoneticians, like other linguists, are most interested in trying to identify those properties that serve a distinctive function. If a property is determined to do so--to be, in other words, "linguistically relevant"--then this finding in itself constitutes the explanation for its presence or absence in any particular piece of speech behavior. But the interest in such properties is, for the phonetician, only one side of the coin; the other side comprises those properties that have little or no apparent cue value for the linguistic identification of an utterance, but that nevertheless display regularities of occurrence that prevent our dismissing them out of hand as simply "noise in the channel." In the case of these latter properties the phonetician is also interested in devising explanations, and here explanations are in a sense much more interesting: whereas the distinctive property is explained by its linguistic function, the linguistically irrelevant property must be explained away, and such an enterprise demands a more strenuous exercise of ingenuity. Most often the explanation offered appeals to mechanical or other physiological constraints on the human organism. Temporal phenomena have frequently been the object of this kind of attention, and among these, certain regularities of vowel duration have been an especially favored topic.

Studies of vowel duration have resulted in two well-known and generally accepted formulations: 1) that the duration of the acoustic segment associated with a vowel depends to a significant extent on the degree of opening of the vowel, and 2) that the duration depends also on the nature of a following consonant. For these relations several explanations have been advanced and apparently accepted, all of them reasonable, but all with certain weaknesses when considered within a more general phonetic framework. A consideration of these explanations, both as to their presuppositions and their implications over and above the particular phenomena they were designed to explain, suggests strongly that, ad hoc to begin with, they have yet to be subjected to the critical testing that must precede their inclusion in any well-integrated theory of speech production. Moreover, the fact that the underlying measurement data derive uniformly from speech samples of a narrowly restricted kind, while it does not relieve us of the duty of trying to explain them, does at the same time raise a question about their precise implication for more spontaneous speech behavior.

---

*Paper presented at the winter meeting of the Linguistic Society of America, San Diego, Calif., 28-30 December 1973.

+Also University of Pennsylvania, Philadelphia.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

Data supporting the statement that for English (and let us restrict ourselves to that) vowel duration is related directly to degree of opening have been reported by many workers--House and Fairbanks (1953), Peterson and Lehiste (1960), and Sharf (1962), to name a few.[1] The relation reported has been understood as a mechanical effect due to a temporal constraint on the movement of the relatively large mass of the lower jaw, with that of the tongue sometimes also implicated: if open or low vowels involve more jaw movement than do the close vowels, then the greater so-called "intrinsic duration" of the former is a natural consequence, provided we believe that in speech we regularly operate close to the limits set by the physical constraints on the mechanism. Lehiste, in her 1970 review of the literature on vowel duration, says very bluntly that "The greater length of low vowels is due to the greater extent of the articulatory movements involved in their production" (p. 19). If we can take the frequency of the first formant as a reasonably good acoustic index of vowel opening, we can see just how closely duration and opening are related.

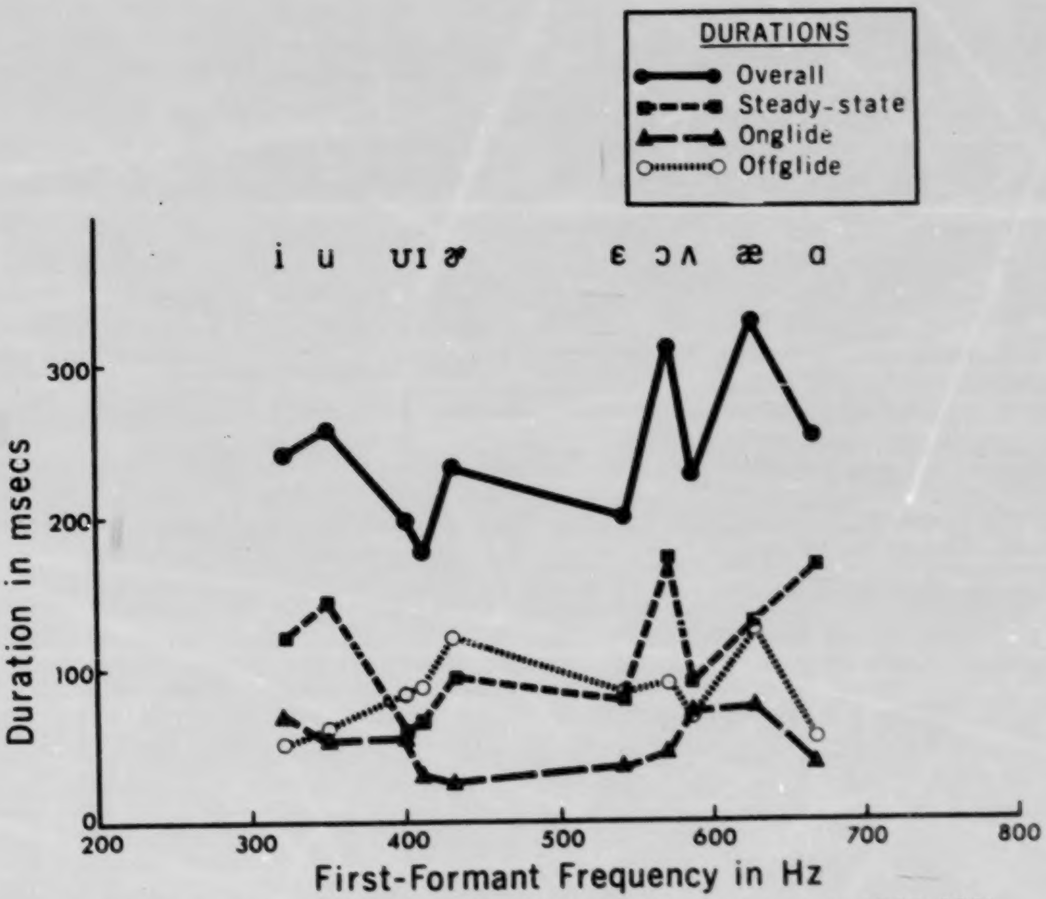
In Figure 1 the topmost line represents mean vowel durations reported by Peterson and Lehiste (1960) as functions of representative values of first-formant frequencies from Peterson and Barney (1952). If the short and long vowels are taken separately, we can, I think, see a tendency for duration to increase with increasing first-formant frequency; at least [I] and [U] are shorter on the average than [ʌ], and [i] and [u] are likewise shorter than [ɔ] and [æ]. The picture is spoiled a bit by [a], which is no longer than [u], but other studies, that of Sharf's (1962) for example, show [a] longer than [u]. Note that the relation of [æ] to [a] is consistent with Perkell's (1969) X-ray finding that although the tongue is higher for [æ] than for [a] the mandible is lower for the former.

The data represented in the three lower lines of Figure 1 are more troublesome. If we are to suppose that the low vowels are longer simply because the mandible is moving as fast as it can, but that it can't manage to cover the required distance in as short a time for those vowels as it does for the high vowels, then we should expect the low vowels to have longer glides and shorter steady-state intervals than the high vowels. But the data from Lehiste and Peterson (1961) show instead the absence of any systematic difference in glide durations for low as against high vowels, and show quite clearly that the greater overall duration of the low vowels, or at any rate of [ɔ] and [a], is primarily a greater duration of their steady-state intervals. In fact the average onglide duration for the vowel [i] is given as 67 msec, while that for [æ] is 73 msec, with [a] only 36 msec. By contrast the steady-state intervals are reported as 120, 132, and 169 msec respectively. It is difficult to see just how a mechanical constraint operates to yield a 90 msec difference in the overall durations of [i] and [æ], while at the same time the latter vowel is on the average produced with a steady-state interval lasting as long as 132 msec. Probing a bit further we find that the vowel for which the sum of the on- and offglide durations is least is [a], while it is greatest for [æ]. These two vowels can hardly be opposed along the vowel height dimension. But the failure to find high vowels with

---

[1] This paper was prepared before the publication of Dennis Klatt's short paper in the Journal of the Acoustical Society of America (54, 1102-1104, 1973), "Interaction between two factors that influence vowel duration," in which data that agree substantially with earlier studies are reported.

VOWEL DURATION vs FIRST-FORMANT FREQUENCY



Vowel durations (mean of 5 talkers) from Peterson-Lehiste 1960, 1961.
First-formant frequencies from Lehiste-Peterson 1961.

FIGURE 1

regularly shorter glides than low vowels is consistent with findings reported by a number of investigators, most recently by Sussman, MacNeilage, and Hanson (1973), that show the velocity of jaw movement to vary directly with its total displacement in vowel-stop and stop-vowel sequences.

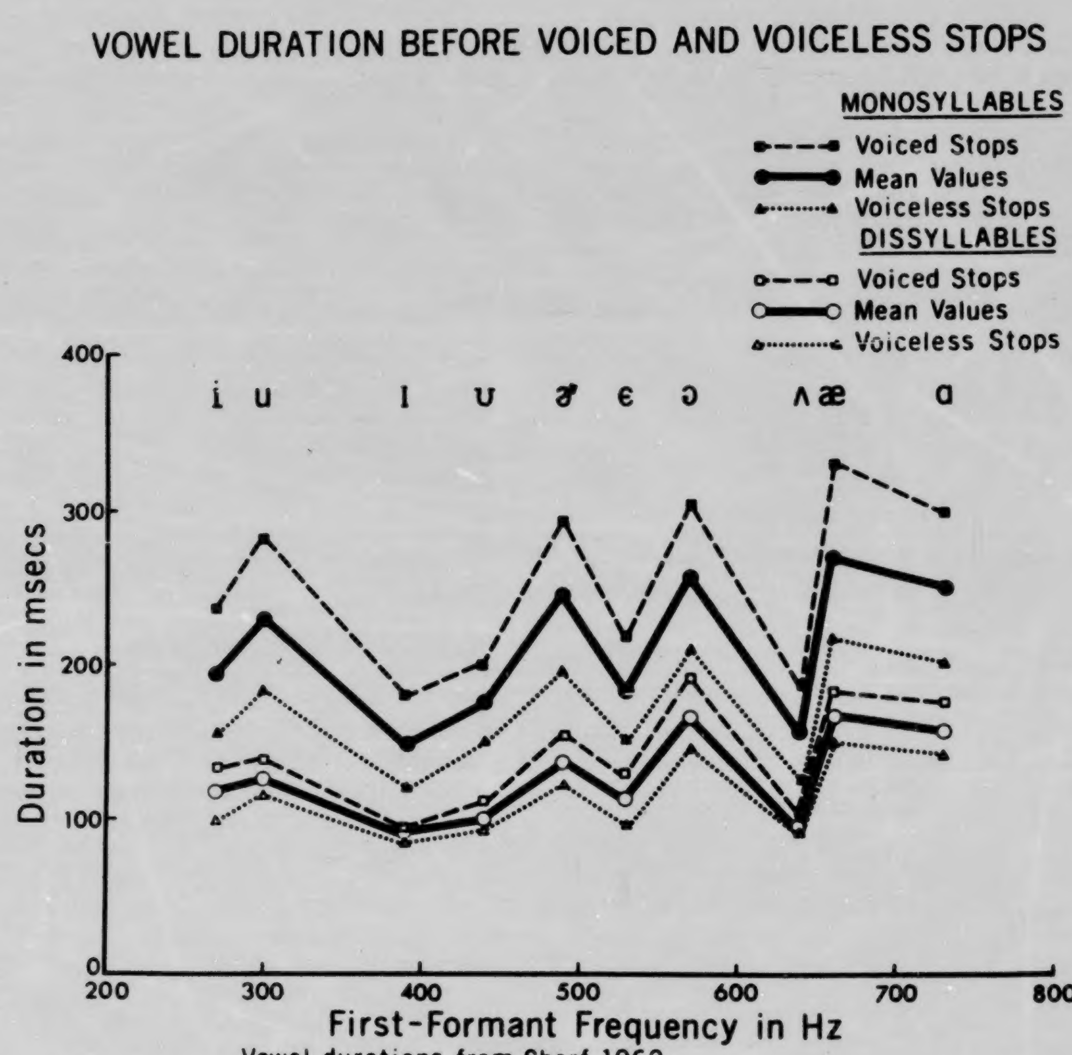
This finding, taken together with the capacity of the system for "target undershoot" and articulatory compensation that Lindblom (1967) has talked about, provides a basis for expecting little or no systematic difference in the durations of high and low vowels. If there is in fact such a difference, it remains nevertheless difficult to maintain the necessity for this difference because of a mechanical constraint on tongue and jaw velocities, given the long steady-state intervals reported by Lehiste and Peterson (1961). Without the Lehiste-Peterson analysis of vowels into glide and steady-state subsegments the explanation of duration variation would be unquestioned; with their analysis, and with the studies on articulatory velocities, it seems to me untenable. One might perhaps better assert that the low vowels, produced with more movement of jaw and perhaps tongue as well, require for perception of the intended vowel that the formant pattern be maintained in a target region over a longer interval "because of" the more extensive formant shifts during the onset and offset glide intervals. But it is dangerous to talk too early of the necessity for any feature. Perhaps it is not a matter of perceptual need at all, but only that since a stronger action is required for the low vowels it is not totally unexpected, on the basis of other phonetic facts, to find that the articulators are maintained longer on target.

The notion that the longer vowels are necessarily longer due to the mechanical inertia of jaw or jaw and tongue runs into other difficulties when one considers vowel duration variation ascribable to differences of context. Figure 2 shows data from Sharf (1962) that are very much in agreement with those of Lehiste and Peterson (1961). (Observe that [a] here is one of the longer vowels.) What is remarkable is how the durational relation among the various vowels is maintained, whether the vowels are in one- or two-syllable words, and whether they are followed by voiced or voiceless stops. The magnitude of these context effects is at least as great as that ascribed to degree of opening. Thus, for example, if it is maintained that [æ] is longer than [I] because of the inertia of the mandible, this constraint is nevertheless suspended if to the monosyllabic word containing [æ] a second syllable is added, or if instead of a following voiced stop there is a voiceless one.

On the general subject of duration in relation to number of syllables there has not, to my knowledge, been any attempt to fashion an explanation, and I will pass on to the relation between vowel duration and the nature of the following consonant, for which the literature provides no fewer than four explanations. These are:

1) Vowels are shorter before voiceless consonants because those consonants are fortes, and fortisness involves the earlier onset of articulatory closure.

2) Vowels before voiceless consonants are shorter because the strong closure gesture is accomplished more rapidly, again because of the fortis nature of those consonants.


VOWEL DURATION BEFORE VOICED AND VOICELESS STOPS

Vowel durations from Sharf 1962
First-formant frequencies from Peterson-Barney 1952

FIGURE 2

3) Vowels are lengthened before voiced stops to allow time for laryngeal readjustment needed if voicing is to be maintained during oral closure.

4) Vowels are longer before voiced and shorter before voiceless consonants according to a rule of constant energy expenditure for the syllable, longer vowels and voiceless consonants both being more costly in articulatory energy.

Of these explanations the last is flawed by the absence of an agreed-upon measure of overall articulatory energy, as well as any rationale for supposing that constancy of energy expenditure, even if measurable, should characterize only one class of sequences, those consisting of a syllabic nucleus and following consonant or consonant cluster. Explanation 3 has, I suppose, the greatest currency at the moment, mainly because of The Sound Pattern of English (1968). As a serious explanation of the relation between vowel duration and following stop it suffers on several counts. First, the available electromyographic and fiberoptic data provide little indication of laryngeal change before voiced stops, but they do indicate that the arytenoid cartilages are subject to an adjustment in rough synchrony with the closure for voiceless stop production. Moreover, the hypothesis that lengthening is required for the maintenance of glottal pulsing throughout the interval of oral closure should be tested, not by asking how much longer is the vowel before voiced than before voiceless stops, but rather how much longer is the vowel before voiced stops than before nasal consonants. The answer to the second question is that the incremental duration is essentially zero. If it is still insisted that vowel lengthening is required to allow time for glottal readjustment before voiced stop closure, then this would seem to imply that the shorter vowels ought to show a greater increase in duration than the longer vowels. But in fact, from Sharf's (1962) data (Figure 3) it appears that the longer a vowel is preceding voiceless stops, the greater the durational increment added before voiced stops. According to Sharf's data the relation between duration before voiceless stops and the durational increment appears to be linear.

Explanations 1 and 2, both of which postulate a shortening before the voiceless consonants because of their claimed greater force of articulation, find confirmation in electromyographic and velocity measurement data that show voiceless stop closures to begin earlier and to be executed more rapidly than closures for the voiced stops. It is odd, however, that this advancement in the timing of closure can be said to be explained by the fortisness feature. This may reflect the prejudice against giving first place to that feature difference between voiced and voiceless stops for which the evidence is incontrovertible--namely the difference in laryngeal state. The assumption is made, and there is no serious attempt made to justify it, that the articulatory program for a consonant involving oral occlusion is independent of whether the stop is voiceless, oral and voiced, or nasalized and voiced. Since the programs for voiced and voiceless stops are clearly not the same when they follow a stressed vowel, this failure must be explained. But the explanation based on the assumed fortisness of the voiceless stops says no more than that the voiceless stops are produced with an earlier and more rapid closure than the voiced ones. The concomitant laryngeal change, abduction of the arytenoids, is tacitly taken to be secondary to the supraglottal event. However, one might just as reasonably suppose that the laryngeal gesture determines the timing of the closure, and that there is no ground for assuming a priori that the onset of arytenoid abduction should follow a temporal program identical with the one that determines the time of closure for the

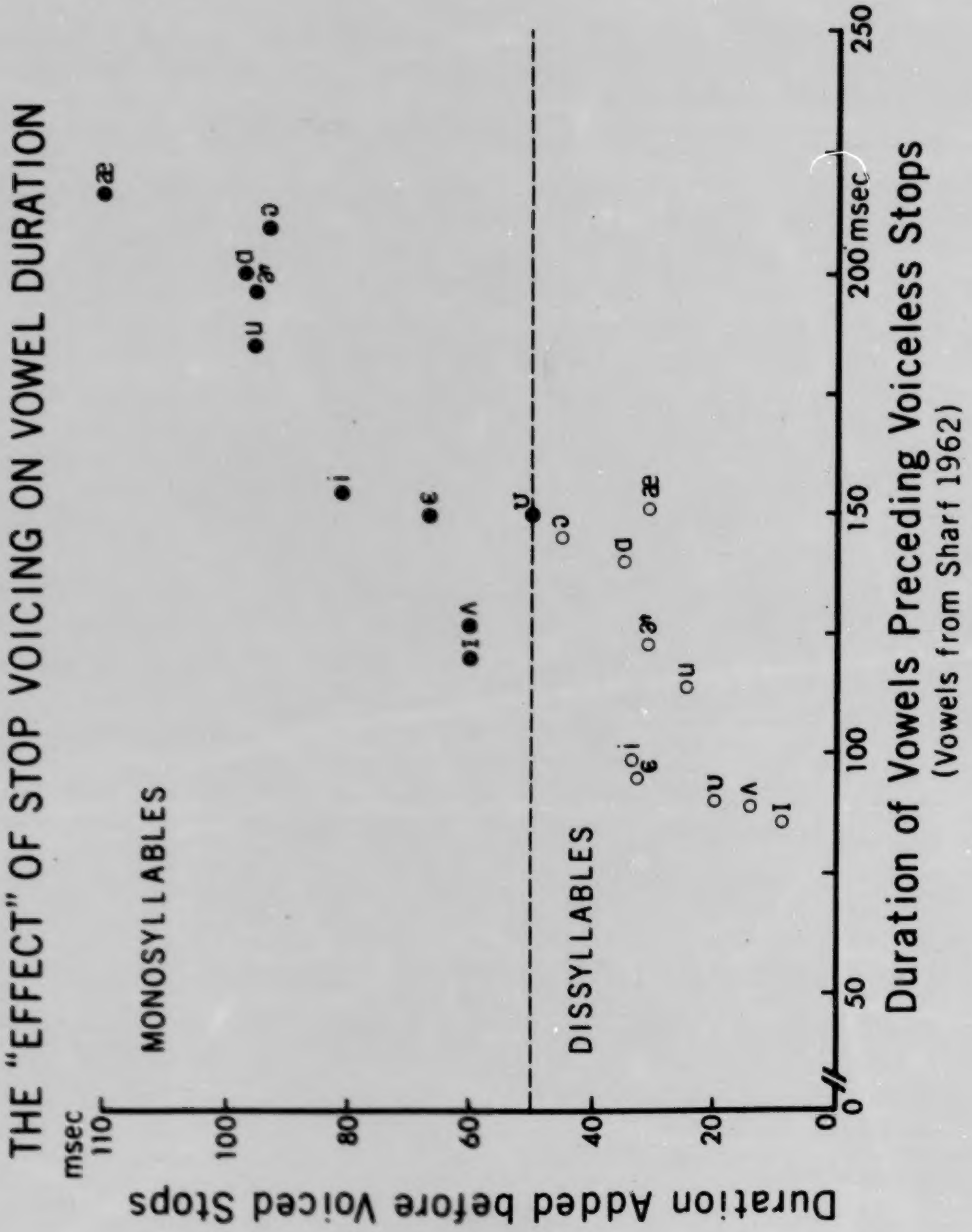
THE "EFFECT" OF STOP VOICING ON VOWEL DURATION
MONOSYLLABLES
DISSYLLABLES
Duration Added before Voiced Stops
msec
110
100
80
60
40
20
0
0
50
100
150
200 msec
250
Duration of Vowels Preceding Voiceless Stops
(Vowels from Sharf 1962)
æ
ɔ
a
e
u
i
ɛ
ʌ
ɪ
ʊ

FIGURE 3

voiced stop. However, there is a reason that can be advanced as to why the closure for the voiceless stop should occur not long after the devoicing gesture begins: that is, that the phonetic result would otherwise be not a sequence of vowel + voiceless stop, but rather, vowel + aspiration + voiceless stop--a phonetic output unacceptable as normal English. Why such an output is unacceptable is a question that the phonetician is not in a position to answer.

REFERENCES

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper and Row).

House, A. S. and G. Fairbanks. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels. J. Acoust. Soc. Amer. 25, 105-113.

Lehiste, I. (1970) Suprasegmentals. (Cambridge, Mass.: MIT Press).

Lehiste, I. and G. E. Peterson. (1961) Transitions, glides, and diphthongs. J. Acoust. Soc. Amer. 33, 268-277.

Lindblom, B. E. F. (1967) Vowel duration and a model of lip mandible coordination. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR 4/1967.

Perkell, J. (1969) Physiology of speech production: Results and implications of a quantitative cineradiographic study. (Cambridge, Mass.: MIT Press).

Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of the vowels. J. Acoust. Soc. Amer. 24, 175-184.

Peterson, G. E. and I. Lehiste. (1960) Duration of syllable nuclei in English. J. Acoust. Soc. Amer. 32, 693-703.

Sharf, D. (1962) Duration of post-stress intervocalic stops and preceding vowels. Lang. Speech 5, 26-30.

Sussman, H. M., P. F. MacNeilage, and R. J. Hanson. (1973) Labial and mandibular movement dynamics during the production of bilabial stop consonants: Preliminary observations. J. Speech Hearing Res. 16, 397-420.

Two Processes in Vowel Recognition: Inferences from Studies of Backward Masking*

M. F. Dorman,[+] D. Kewley-Port,[++] S. Brady,[+++] and M. T. Turvey[+++]
Haskins Laboratories, New Haven, Conn.

INTRODUCTION

From an information-processing approach (Haber, 1969a), perception can be viewed as a hierarchically organized set of operations that extract over time different categories of information from a signal. In studies of visual perception, backward recognition masking of form by a patterned stimulus has become a powerful tool for probing stages of perceptual processing (Haber, 1969b; Turvey, 1973). The use of this technique is based on the assumption that when a masking stimulus follows a target stimulus after some delay, processing of the target occurs during the delay, but further processing is either distorted or interrupted by the arrival of the mask.

Although peripheral and central operations in vision have been investigated by systematic variation of the physical and temporal parameters of target and mask stimuli in forward and backward masking paradigms (Turvey, 1973), few studies have used these techniques to probe the recognition of vowels. Massaro (1972b, 1974) presented listeners with 20 msec vowels, either /i/ or /I/, followed at interstimulus intervals (ISIs) from 0-500 msec by a 270 msec mask (alternating segments of /a/ and /u/). Vowel recognition was near chance at 0 msec ISI and reached asymptote by 250 msec ISI. From these data and from studies of backward masking of nonspeech auditory signals (Massaro, 1972a), Massaro concluded that perceptual processing time for auditory signals can last between 120 and 250 msec. Consequently, Massaro has suggested that the range of vowel durations found in normal speech [150-350 msec (House, 1961)] may be constrained by the necessity to evade backward masking by following segments of the speech signal. A somewhat different conclusion can be drawn from Pisoni's (1972) studies of backward masking of vowels. Listeners were presented computer-generated 40 msec vowels /i, I, ε/ followed by another vowel from the same set. Performance at 0 msec ISI was 85 percent correct and reached asymptote by 80 msec ISI. These data suggest that when vowels are long enough to be readily identified in isolation, minimal backward masking is obtained.

---

*Expanded version of a paper presented at the 87th meeting of the Acoustical Society of America, New York, April 1974.

[+] Also Lehman College of the City University of New York.

[++] Also the Graduate Center of the City University of New York.

[+++] Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

While the studies cited above serve as a useful beginning, Turvey (1973) has pointed out that inferences about perceptual processing from studies employing masking techniques can best be made only after systematic variation of the relationship between target and mask stimuli. For example, in studies of visual backward masking, estimates of perceptual processing time vary as a function of the stimulus used as a mask [a random noise mask presented monoptically disrupts target recognition over a shorter interval than a patterned mask (Turvey, 1973)].

The present paper reports four experiments in which selected parameters of target and mask stimuli were varied in forward and backward masking tasks to investigate perceptual operations in the recognition of brief vowels.

EXPERIMENT I

The purpose of Experiment I was to determine the vowel durations over which masking occurred. A long series of informal listening tests suggested that only very brief real-speech vowels suffered backward masking. Therefore, vowel sets of three durations (15.5, 20, and 30 msec) were constructed and presented to listeners in both forward and backward masking tasks.

Method

Subjects. The *S*s were undergraduate students from Yale University and the University of Connecticut. Yale University students received $2.00 per hour for participation; University of Connecticut students received class credit.

Preparation of stimuli. The target stimuli were the vowels /i/, /ɛ/, and /ʌ/ spoken in isolation by a male with a fundamental frequency of approximately 120 Hz. Using the Haskins Laboratories computer-controlled PCM system (Cooper and Mattingly, 1969) three sets of vowels were prepared. Segments of 15.5, 20, and 30 msec duration were excised from steady-state portions of each vowel. The mask was a computer-synthesized two-formant sound of 125 msec duration with formant frequencies at 489 Hz and 1690 Hz. This mask was vowel-like but did not have formant frequencies similar to any English vowel. The target and mask stimuli were equated for peak-to-peak amplitude. Special care was exercised to insure that the stimuli were recorded at the best possible signal-to-noise ratio, approximately 40 db.

Training materials. Under computer control one sequence of three repetitions of the target vowels and six 18-item sequences of target vowels (six repetitions of each vowel in each randomized sequence) were recorded on audio tape. The intertrial interval was 4 sec for all sequences.

Test materials. Six test sequences were constructed under computer control. For each vowel set duration (15.5, 20, and 30 msec) both a forward and backward masking sequence were generated. (A six-item practice sequence was also generated for each test sequence.) In the forward masking condition the mask preceded the target vowels at intervals of 0, 25, 50, 100, 200, and 500 msec. Each vowel occurred six times at each ISI. In the backward masking condition the vowels preceded the mask at intervals of 0, 25, 50, 100, 200, and 500 msec. Each vowel occurred six times at each ISI. The sequence of vowels and ISIs was randomized in each test sequence. Each test sequence was presented twice, thus creating two blocks of 54 trials, or a test sequence of 108 items.

Apparatus. The stimuli were recorded and reproduced on an Ampex AG500 tape recorder. The tape recorder output was interfaced with a distribution amplifier which insured equal signal amplification into four sets of matched Grason-Stadler TDH39-300Z earphones. The stimuli were presented diotically at a comfortable listening level. A calibration signal insured equal signal levels in all conditions within and between experiments.

Design. One group of *S*s was trained and tested with 15.5 msec vowels, another group with 20 msec vowels, and a third group with 30 msec vowels. All *S*s were tested on both the forward and backward masking sequences in counterbalanced order.

Procedure

The *S*s were seated in a large sound-attenuated room and were told they would hear three very brief vowels, /i/, /ɛ/, and /ʌ/, which they were to learn to identify. First, the *S*s were presented three repetitions of the three-vowel set. Next, the *S*s were told they would hear six 18-item lists of the vowels in random order and were instructed to write the identity of the vowels on printed response sheets. The correct responses, initially covered by a movable slider, were printed next to the space for the *S*s' responses. By moving the slider down the page for each succeeding trial the *S*s uncovered the correct responses for the preceding trial, thus providing immediate feedback of correct responses. On the final 18-item sequence, the *S*s were given no feedback of correct responses.

After a brief rest period, the *S*s were told they would hear, in one sequence, the vowels followed by a mask at various intervals, and in another sequence, the mask followed by the vowels at various intervals. The *S*s were instructed to write the identity of the vowels on a printed answer sheet. After six practice trials the *S*s were presented a 108-item test sequence in two blocks of 54 trials. Then, after a brief rest, the *S*s were given another six practice trials and the other 108-item test sequence.

Results

Only those *S*s who made no errors on the final (no feedback) practice sequence were considered in the data analyses.[1] All 10 *S*s trained with the 30 msec vowels achieved perfect performance on the final practice sequence. Of the *S*s trained with the 20 msec vowels, 83 percent achieved perfect performance, and with the 15.5 msec vowels, 62 percent achieved perfect performance. Clearly, identification of the 20 and 30 msec vowels in isolation was a relatively easy task.

Inspection of errors as a function of ISI in the backward masking condition revealed two distinct error patterns. One population was characterized by better than 80 percent correct responses at the 0 msec ISI and very few errors at other ISIs. These *S*s will be referred to as Nonmaskers. Another population was characterized by scores of less than 80 percent correct at 0 msec ISI while not reaching asymptotic performance until 100-200 msec ISI. These Ss will be referred to as Maskers.

---

[1] Eight *S*s who made errors (average = 79 percent correct) on the final training test with 15.5 msec vowels were tested in the backward masking sequences. For these *S*s vowel recognition did not markedly improve with an increase in ISI. At 0 msec ISI vowel recognition was 55 percent correct; at 500 msec ISI, 66 percent correct.

The results for the 15.5, 20, and 30 msec vowel duration groups in the backward masking condition are shown in Figure 1. None of the Ss tested with the 30 msec stimuli were classified as Maskers. Seven of the ten Ss in the 20 msec condition were classified as Nonmaskers, and three as Maskers. Of the ten Ss in the 15.5 msec group, six were classified as Nonmaskers and four as Maskers on the test sequence. (The depressed score for the Maskers at 50 msec ISI reflected the poor performance of only one S.)

In the forward masking condition, all of the groups achieved 97 percent or better correct responses at all of the ISIs.

Discussion

No forward masking was observed in any of the conditions. Since the 15.5 msec vowels were of minimum duration, it appears that perceptual interference in the recognition of vowels occurs only when a masking stimulus follows a target stimulus.

For the backward masking task, in the 30 msec vowel condition at 0 msec ISI, recognition of vowel targets was essentially perfect. In the 20 msec vowel condition at 0 msec ISI, the majority of Ss showed very little or no impairment in vowel target recognition. Even when the vowels contained only one complete pitch period, the majority of Ss performed at better than 90 percent accuracy at 0 msec ISI. From these data we conclude that for the majority of Ss, a stimulus duration of between 20 and 30 msec is sufficient for processing mechanisms to separate the target vowels from the mask and to extract the features necessary for the recognition of the vowels. Of course, this conclusion applies only to the specific conditions of Experiment I (i.e., a three-vowel target set and a 125 msec vowel-like mask).

To determine whether the outcome of Experiment I was a function of the duration of the stimuli, or of the number of pitch pulses in the stimuli, a series of 15.5 and 20 msec stimuli were created from vowels spoken by a female with a fundamental frequency of approximately 250 Hz. The 15.5 msec vowels contained three complete pitch pulses; the 20 msec vowels contained five pitch pulses. The 15.5 and 20 msec vowels were much more difficult to recognize in isolation than the male vowels. Of the Ss trained with the 20 msec stimuli, only 50 percent achieved perfect performance on the final practice list; for the 15.5 msec vowels only 16 percent achieved perfect performance. It is possible that the 3.5 kHz digital filtering on the PCM system may have degraded these vowels.

The design of the experiment was identical to that of Experiment I. The results are shown in Figure 2. On the backward masking sequences two types of Ss were again identified, Nonmaskers and Maskers. The proportion of Nonmaskers to Maskers was similar to that in the male-vowel condition. Clearly there was no large decrease in the amount of masking with the increase in the number of pitch pulses in the signal. It appears that the masking observed for brief vowels results from the brevity of the signal and not from the number of pitch pulses in the signal. It is important to note that the near perfect vowel recognition at 0 msec ISI, for at least the majority of Ss with the 15.5 msec vowels, was not

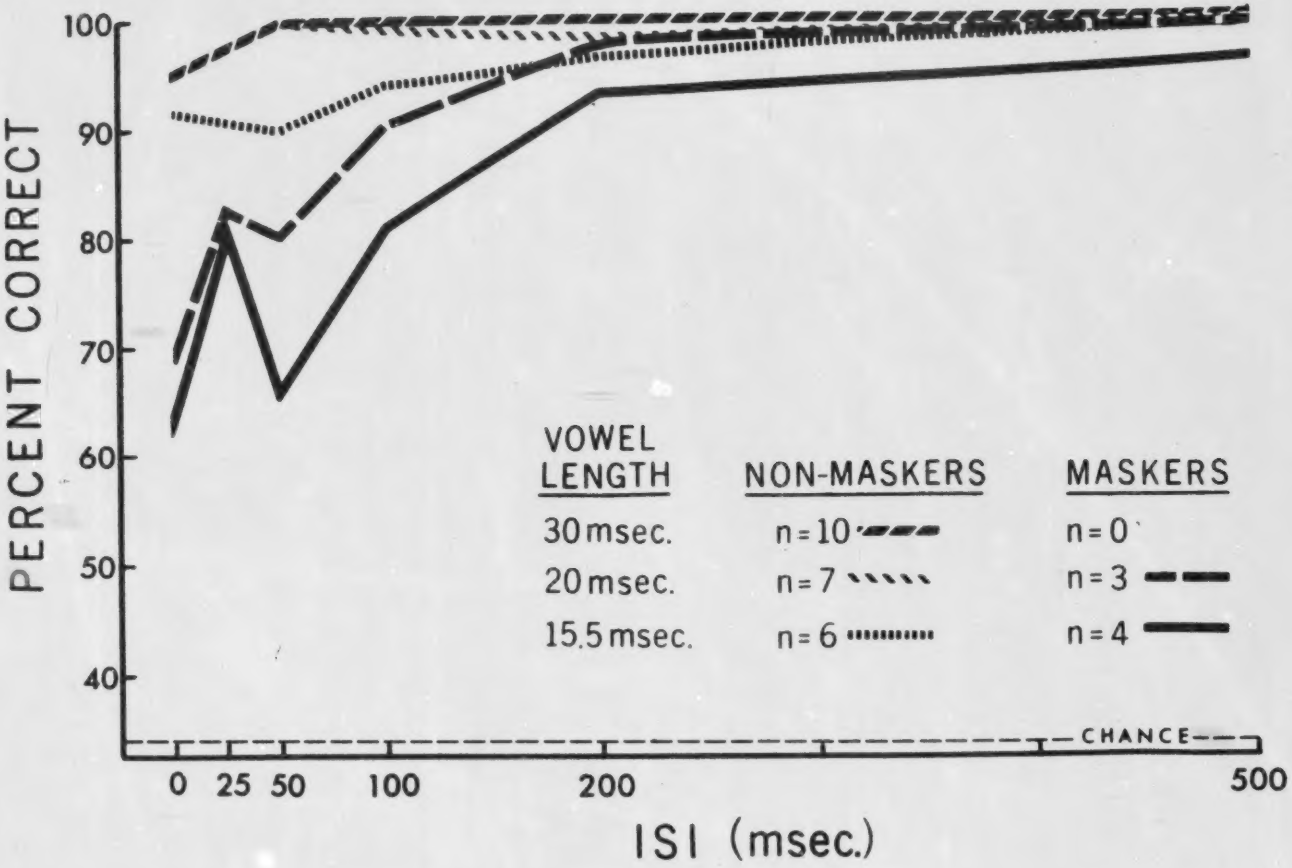


FIGURE 1

Figure 1: Average percent correct male-vowel identification as a function of vowel duration and interstimulus interval.

FIGURE 2

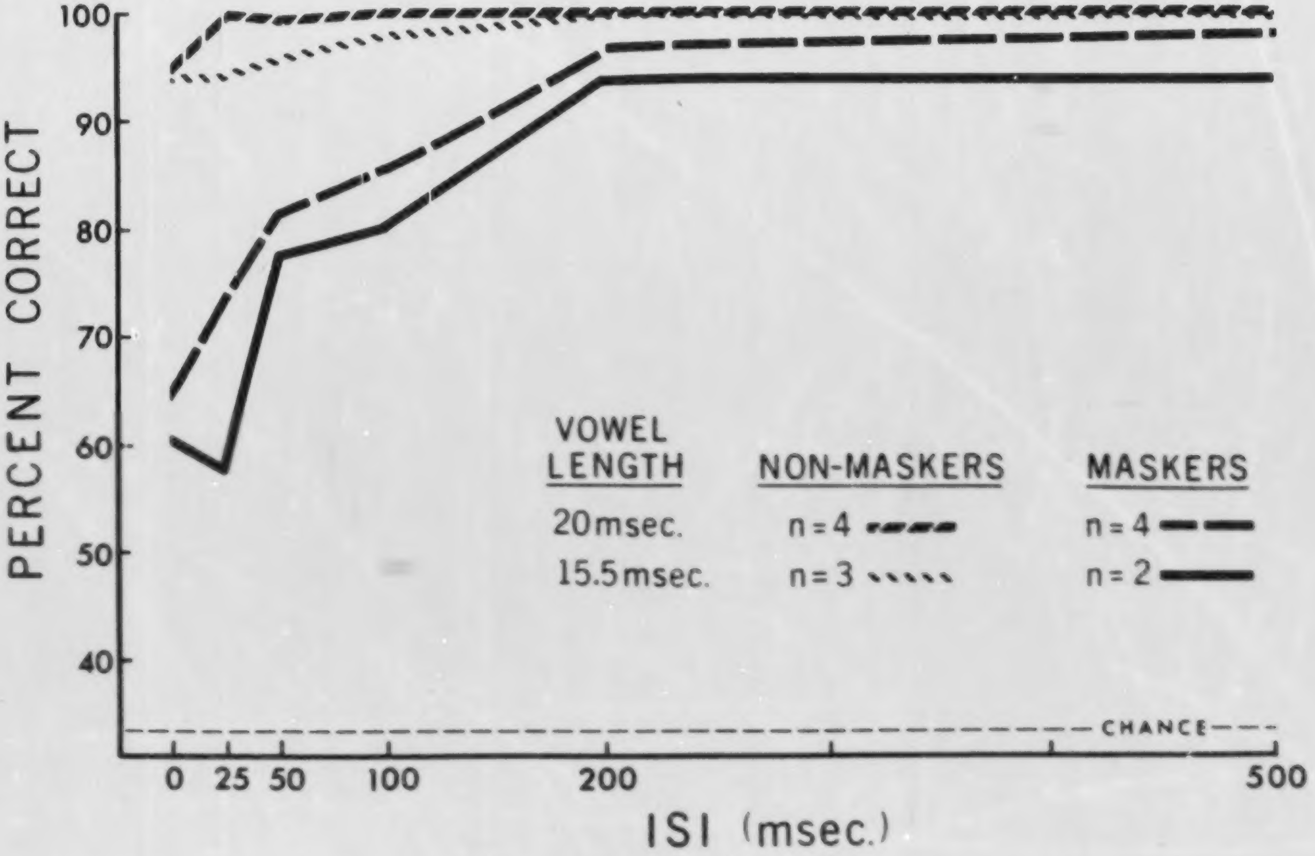


Figure 2: Average percent correct female-vowel identification as a function of vowel duration and interstimulus interval.

simply a function of an "easy" discrimination (cf. Massaro, 1972b:134), since, as noted above, only 16 percent of the Ss tested could identify the vowels in isolation.

In the 15.5 and 20 msec vowel conditions, 39 percent of the Ss (averaged over both male- and female-vowel conditions) were characterized as Maskers, i.e., vowel recognition at 0 msec ISI was approximately 65 percent correct and did not reach asymptote until 100-200 msec ISI. There are several possible explanations for the observed backward masking. Following Kahneman (1968) and Turvey (1973), we can initially distinguish between peripheral (sensory) and central loci of interference and between two mechanisms of interference--peripheral integration of target and mask, and central interruption of target categorization by the mask. The integration hypothesis assumes that the target and mask interact peripherally, thus presenting a "noisy" representation to the central processor for recognition. This hypothesis predicts interference in target recognition in both forward and backward masking paradigms.

The interruption hypothesis suggests that a clear target representation arrives at the central processors but categorization of the target representation is disrupted by the arrival of the mask before recognition is achieved. This hypothesis predicts severe impairment of target recognition only in backward masking paradigms. Since only backward masking was observed for the Maskers, we infer that the locus of interference was central and resulted from the mask disrupting the categorization of the target representation.

The long ISI necessary to evade masking for the Maskers in the 15.5 msec and 20 msec vowel conditions is in marked contrast to the essentially perfect performance of all Ss at 0 msec ISI in the 30 msec vowel condition. These data suggest that silent processing time and stimulus duration do not have additive effects in determining vowel recognition. Allowing a 25 or 50 msec silent processing interval after the 15 and 20 msec targets was not equivalent in terms of facilitating vowel recognition to adding 10 msec stimulus duration. (For another discussion of the effects of silent processing time and stimulus duration in the recognition of vowels, see Massaro, 1974.)

EXPERIMENT II

Experiment I investigated the effect of several target parameters on vowel masking. Experiment II varied two mask parameters. In visual backward masking, impairment of target recognition varies as a function of the similarity of target and mask features (Schiller, 1965; Kahneman, 1968). Since the mask of Experiment I was a synthesized two-formant vowel-like stimulus, it is possible that a mask that shared more features with the target vowels would produce more interference with vowel recognition. To investigate this, in one condition of Experiment II the mask was a 125 msec vowel /o/. Turvey (1973) has shown for central backward masking in vision that an increase in mask energy beyond that of target energy does not increase the extent of backward masking. To determine whether mask energy affects vowel recognition, in another condition of Experiment II the mask was the two-formant mask of Experiment I increased in intensity 20 db.

Method

Subjects. The Ss were undergraduate students from Yale University and the University of Connecticut. Yale University students received $2.00 per hour for participation; University of Connecticut students received class credit.

Preparation of stimuli. The target vowels were the 20 msec vowels used in Experiment I. Two masks were constructed. One mask was the vowel /o/, spoken by the same male speaker as in Experiment I, truncated to 125 msec duration, and equated for peak-to-peak amplitude with the target vowels. A second mask was the two-formant mask of Experiment I, but made 20 db (true RMS) more intense than the mask of Experiment I.

Test materials. Backward and forward masking sequences were generated with both the /o/ mask and the +20 db two-formant mask. The internal construction of the test sequences was the same as in Experiment I.

Design. One group of Ss was tested with the /o/ mask in both the forward and backward masking conditions in counterbalanced order. Another group of Ss was tested with the +20 db mask in the forward and backward masking conditions in counterbalanced order.

The Training materials, Apparatus, and Procedure were the same as in Experiment I.

Results

Only those Ss who made no errors on the final practice sequence were considered in the data analyses. Of the Ss trained for the /o/ mask condition, 83 percent achieved perfect performance on the final practice sequence; of the Ss trained for the +20 db mask condition, 90 percent achieved perfect performance.

The results for the Maskers in the /o/ mask and +20 db mask conditions and the results from the Maskers in the 20 msec vowel condition of Experiment I are shown in Figure 3. Five of the ten Ss in the /o/ mask condition were characterized as Nonmaskers, and five as Maskers. Eight Ss in the +20 db mask condition were characterized as Nonmaskers, and two as Maskers.

In the forward masking conditions, both groups of Ss achieved 92 percent or better correct responses at all of the ISIs.

Discussion

The absence of forward masking in either the /o/ mask or +20 db mask condition reinforces the impression gained from Experiment I that perceptual interference in the recognition of vowels occurs only when a mask follows a target stimulus.

In the backward masking sequences, the performance of the Ss did not differ greatly from that of the Ss in the 20 msec vowel two-formant mask condition of Experiment I. One group of Ss was characterized as Nonmaskers and a smaller group as Maskers. Increasing mask energy did not increase the number of Maskers or increase the ISI necessary to evade masking. The /o/ mask did appear to be somewhat more effective than either of the two-formant masks in terms of the number of Maskers, and in terms of percent correct vowel recognition at 0 msec ISI. Overall, however, the differences between the groups were small. Viewed as a whole, the results from Experiments I and II, i.e., the complete absence of forward masking and the absence of an increase in backward masking with a large increase in mask energy, reinforce the conclusion that the locus of interference for the Maskers was of central rather than peripheral or sensory origin (cf.

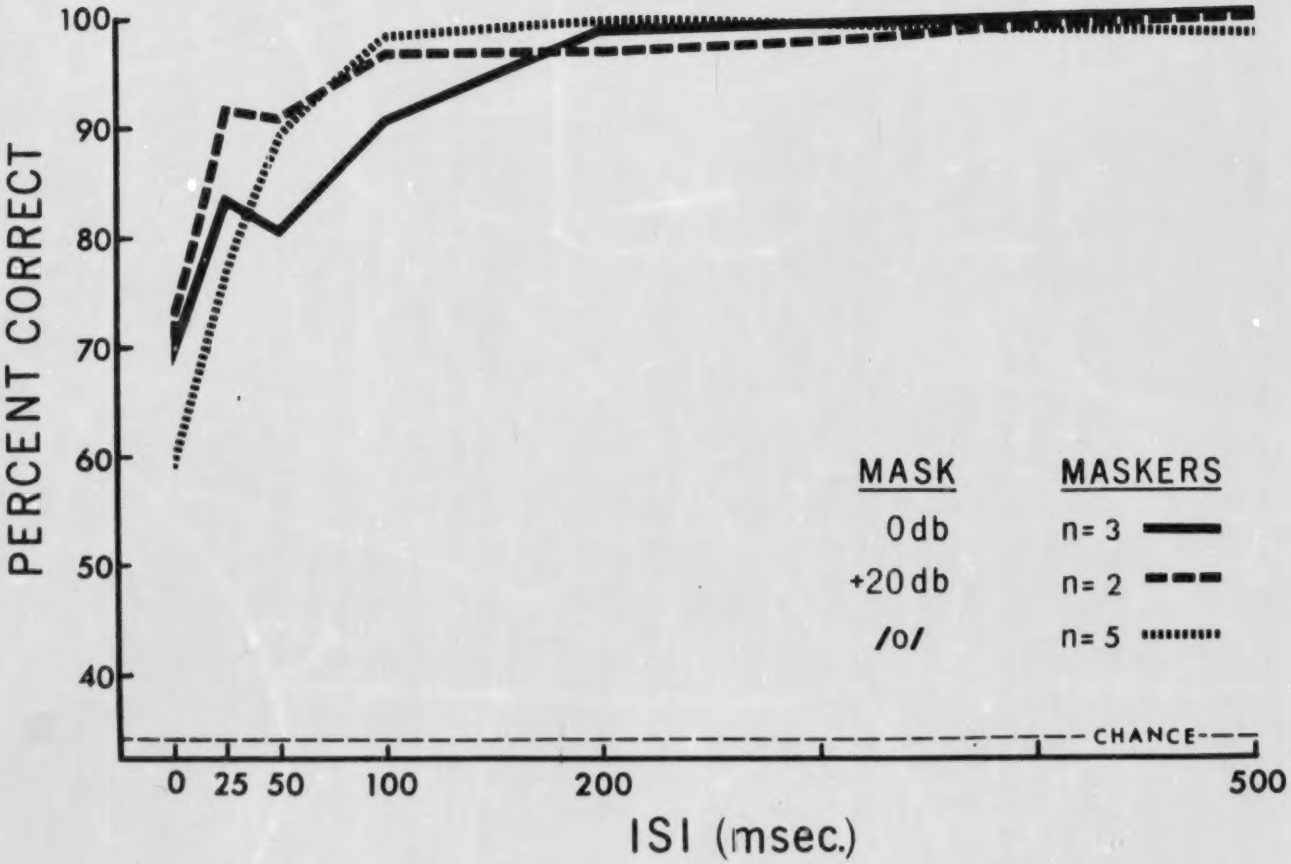


FIGURE 3

Figure 3: Average percent correct vowel identification as a function of masking stimulus and interstimulus interval.

Turvey, 1973:36). The data do not reveal, however, the nature of the difference in perceptual processing for the Nonmasker and Masker populations.

EXPERIMENT III

The outcome of Experiment II confirmed the presence of two listener populations in the backward masking tasks. The magnitude of the difference in ISI necessary to evade masking for the two groups of listeners suggested that the Nonmaskers and Maskers may have used rather different recognition routines to identify the target vowels.

The Maskers' performance is consistent with "constructive" models of stimulus recognition which require that a series of operations be performed on stimulus information before recognition is achieved (Neisser, 1967; Sternberg, 1967). Common to these models is an initial operation of encoding a stimulus as a set of abstract features, and a second operation of comparing the abstracted stimulus representation with a set of stored features in long-term memory. In this broad view of stimulus recognition, increasing the size of the target set in a backward masking task should increase the latency of target categorization (cf., Sternberg, 1966; Massaro, 1974) and should, therefore, increase the susceptibility of the recognition process to interference from a masking stimulus.

In order to probe possible differences in vowel recognition strategy used by the Nonmaskers and Maskers, in Experiment III vowels from sets of two, three, or four vowels were presented to listeners in a backward masking task.

Method

Preparation of stimuli. The target vowels were the 20 msec vowels of Experiment I with the addition of /I/. The mask was the two-formant stimulus of Experiment I.

Training materials. One sequence of three repetitions of the vowel set [i I ɛ ʌ] and five 24-item sequences of vowels (six repetitions of each vowel in each randomized sequence) were recorded on audio tape.

Test materials. Backward masking sequences were generated for the two-, three-, and four-vowel target sets using the same ISIs as Experiment I.

Design. Each S was tested on the two-, three-, and four-vowel backward masking sequences. One group was tested in the order 2, 3, 4, another group 3, 4, 2, and a third group 4, 2, 3.

Procedure. The training procedures were similar of that used in Experiments I and II, modified for four target vowels.

Results

Only those Ss who made no errors on the final practice sequence were considered in the data analyses. Of the Ss trained, 72 percent achieved perfect performance on the final practice sequence. The Ss were classified as Maskers and Nonmaskers on the basis of performance on the three-vowel target set. Six Ss (two from each test order) were classified as Maskers, and 12 as Nonmaskers. The averaged performance of the Nonmaskers as a function of target set size is shown in Figure 4. The performance of the Maskers is shown in Figure 5.

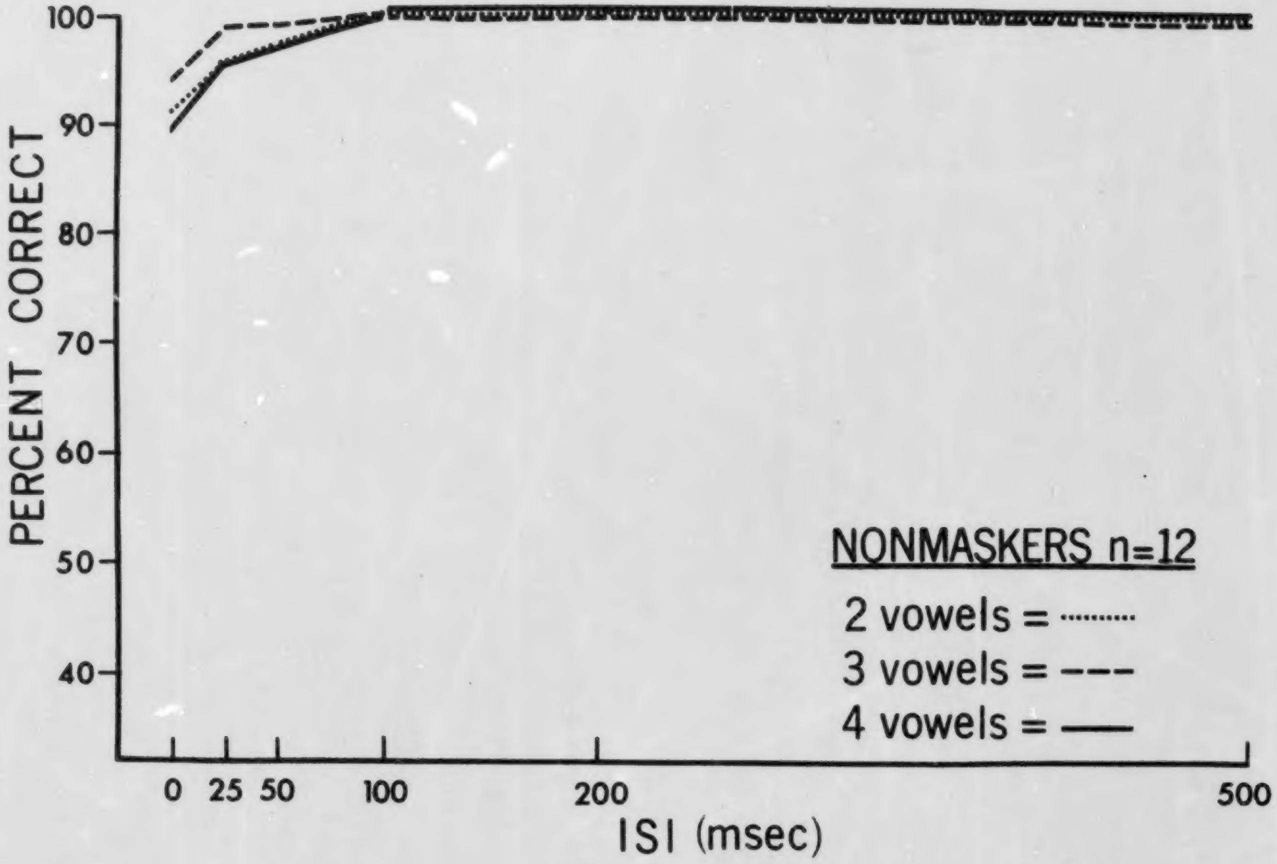


FIGURE 4

Figure 4: Average percent correct vowel identification for Nonmaskers as a function of the number of vowels in the target set and of interstimulus interval.

FIGURE 5

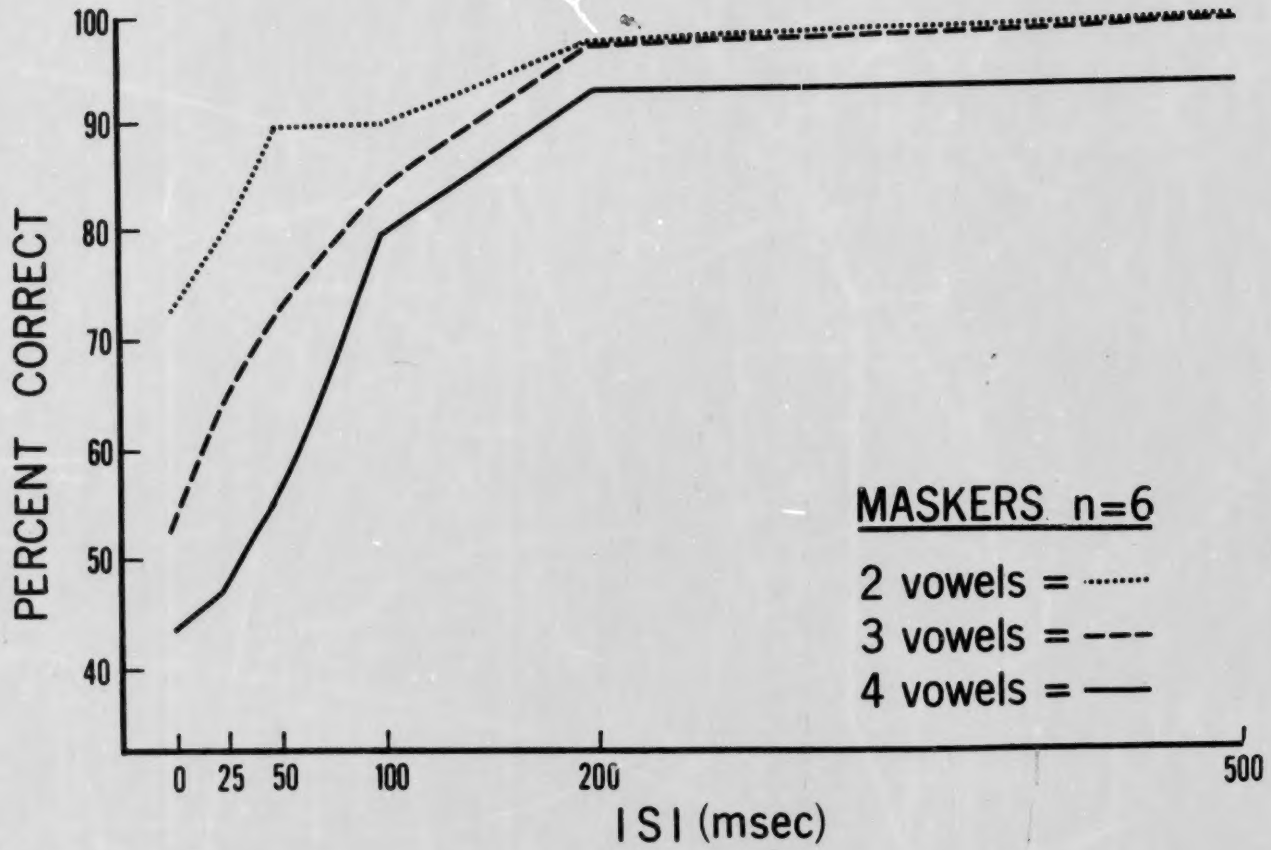


Figure 5: Average percent correct vowel identification for Maskers as a function of the number of vowels in the target set and of interstimulus interval.

The ISI at which target recognition reached asymptote was determined by comparing performance at 500 msec ISI with performance at the shorter ISIs. For the Maskers, on the two-vowel set the first point of difference from 500 msec ISI was at 25 msec ISI ($t_5 = 2.78$, $p < .05$); for the three-vowel set, at 100 msec ISI ($t_5 = 4.11$, $p < .02$); and for the four-vowel set, at 100 msec ISI ($t_5 = 5.46$, $p < .01$). For the Nonmaskers, on the two-vowel set, performance differed from that at 500 msec ISI only at 0 msec ISI ($t_{11} = 2.92$, $p < .02$); for the three-vowel set, at no point; and for the four-vowel set, at 0 msec ISI ($t_{11} = 4.59$, $p < .01$).

Discussion

Two listener populations were again identified in the three-vowel condition. For the Nonmaskers increasing target set size from two to four vowels did not systematically increase backward masking (i.e., performance on the two-vowel set was similar to that on the four-vowel set). For the Maskers, however, performance was systematically affected as a function of target set size. In terms of level of performance at brief ISIs, and in the extent of masking, performance in the two-vowel condition was better than in the four-vowel condition. The absence of a difference in extent of masking for the three- and four-vowel conditions may have been a function of the few data points between 100 and 500 msec ISI.

We should note that the number of vowels in the target sets was at least partially confounded by acoustic similarity between the vowels (e.g., the acoustic difference between /ɛ/-/ʌ/ is greater than that between each member of the set /i/-/I/-/ɛ/). However, it is interesting that the Nonmaskers' performance was unaffected by increasing the number of items in the target set even when the task was complicated by increased acoustic similarity among the items to be recognized. For Maskers, target set size and acoustic similarity between targets could exert independent effects on vowel recognition. This remains to be determined.

EXPERIMENT IV

In Experiments I-III the Nonmaskers were apparently able to determine the identity of the target vowels before the mask disrupted the recognition process. With this interpretation of the Nonmaskers' performance, a change in stimulus parameters that delays or retards the recognition process should make the Nonmaskers more susceptible to backward masking. Degrading the vowel stimuli with white noise should increase the amount of backward masking, as Sternberg (1967) has shown that adding noise to a visual stimulus in a character recognition task increases the latency of encoding a stimulus as a set of abstract features.

Method

Subjects. The Ss were undergraduate students at Yale University who were paid $2.00 per hour for participation.

Preparation of stimuli. For one condition (control) the target stimuli were the 20 msec vowels and 125 msec two-formant mask of Experiment I. For a second condition (noise-added) the target stimuli were constructed by adding white noise 15 db less intense than the vowels to the 20 msec vowels of Experiment I. (Extensive pilot experiments indicated that when greater amounts of noise were added, most naive Ss were not able to identify the vowels.) White noise was similarly added to the mask.

Preparation of test sequences. Backward masking sequences were generated for both the control and noise-added conditions. The internal constraints on the test sequence constructions were the same as in Experiment I. Different stimulus randomizations were used in the two test sequences.

Design. Each S was tested in both the noise-added and control condition. Test order was counterbalanced across Ss.

Procedure. The training procedure was similar to that used in the previous experiments. The Ss were first given three practice sequences with the control vowels, then three practice sequences with the noise-added vowels. Finally, the Ss were given identification tests separately for the control and noise-added vowels.

Results

Only those listeners who made no errors on both the final control and noise-added identification tests were considered in the data analyses. Of the Ss trained, 83 percent achieved perfect performance on the final practice list. The listeners were classified as Nonmaskers and Maskers on the basis of performance in the control condition. In the control condition, ten listeners were classified as Nonmaskers, and two as Maskers. Of the ten Nonmaskers six made more errors in the noise-added condition (Figure 6). For these Ss, performance in the noise-added condition differed from that in the control condition at 0 msec ISI ($t_5 = 4.30$, $p < .01$), at 25 msec ISI ($t_5 = 4.38$, $p < .01$), and at 50 msec ISI ($t_5 = 2.71$, $p < .05$). Four listeners made no errors in either condition. One of these Ss was tested further. No masking was found with -12 db and -9 db noise-added stimuli. Only in a -3 db noise-added condition was masking apparent. The two Maskers made more errors in the noise-added condition than in the control condition (Figure 7).

Discussion

For the majority of listeners who showed no masking of 20 msec vowels at 0 msec ISI, the addition of -15 db white noise to the targets resulted in backward masking extended to 50-100 msec ISI. In terms of an information-processing analysis, the noise in the stimuli could be viewed as increasing the time necessary to extract and encode the vowel features. This increased processing time would, in turn, increase the vulnerability of the recognition routine to disruption by the stimulus arriving second.

The performance of the listeners in the noise-added condition comes closest to the masking functions reported by Massaro (1972b; for a two-vowel discrimination, performance at 0 msec ISI was near chance and reached asymptote by 250 msec ISI). Massaro interpreted his data as providing evidence that the 150-350 msec average vowel duration found in running speech may be necessary to evade backward masking from following segments of the speech signal. However, if Massaro's masking functions were influenced by variables functionally similar to the noise condition of Experiment IV, then such an interpretation would be unduly pessimistic. The absence of masking for the 30 msec vowels of Experiment I and for the majority of listeners in the four-vowel condition of Experiment III suggests that backward masking does not impose a serious constraint on vowel perception or production in running speech.

FIGURE 6

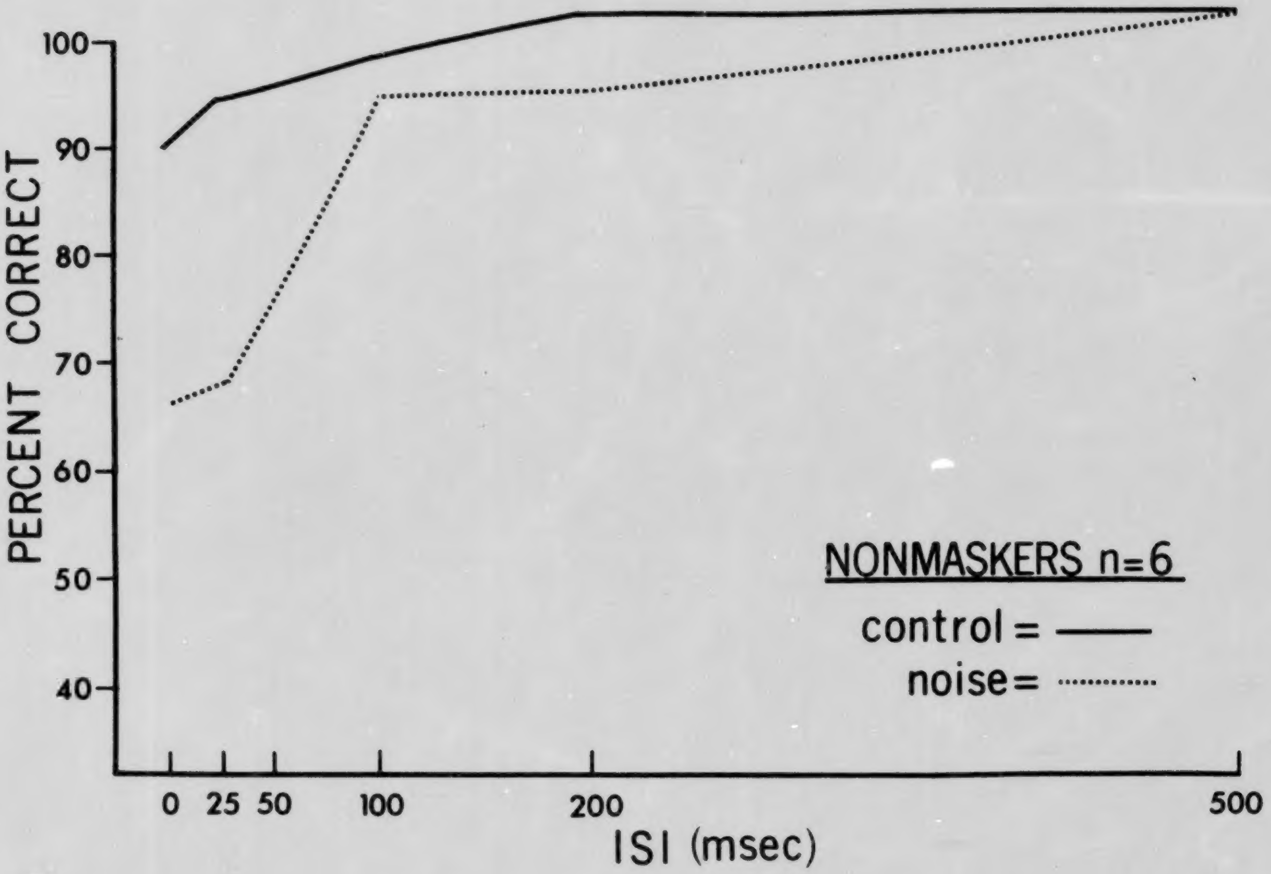


Figure 6: Average percent correct vowel identification for Nonmaskers in the control and noise-added conditions as a function of interstimulus interval.

FIGURE 7

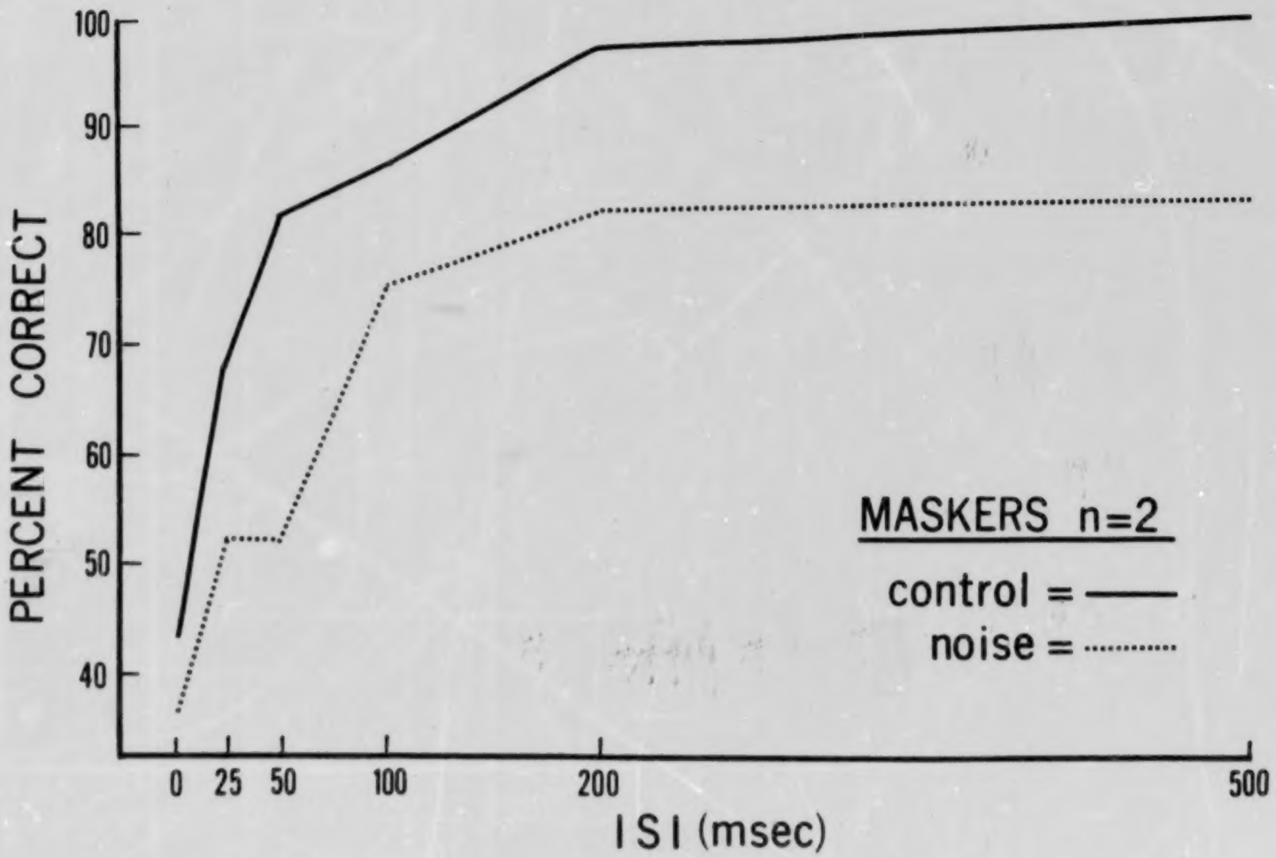


Figure 7: Average percent correct vowel identification for Maskers in the control and noise-added conditions as a function of interstimulus interval.

GENERAL DISCUSSION

One of the central issues generated by the data of Experiments I-IV is the nature of the difference in perceptual processing between the Nonmasker and Masker populations. One possibility is that the two groups simply reflect a quantitative difference in the rate of perceptual processing. In this view, the populations of Nonmaskers and Maskers would rest on opposite ends of a continuum of processing time. Another possibility, however, is that the two groups reflect the outcome of two qualitatively different modes of vowel processing. Depending on the information-processing constraints imposed by a particular experimental task, an individual could employ one or the other of these modes to recognize the vowel targets. This proposal seems to best fit the present data and, indeed, is in keeping with analyses of vowel perception by Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) and Studdert-Kennedy (in press). In the following sections, we review data relevant to this issue.

Perceptual Processing in Speech and Auditory Modes

The relationship between acoustic signal and perceived phoneme, at least for stop consonants, is that of a complex code, i.e., there is a marked lack of invariance between the acoustic signal and phonetic message (Liberman et al., 1967). For this reason stop consonants have been termed "highly encoded" signals. Given the complex relationship between signal and message, it is not surprising that stop consonants appear to be perceived in a mode (the speech mode) different from that of simple nonspeech auditory signals. As Liberman (1973) has pointed out, perception in the speech mode can be characterized as abstract and categorical. Listeners are unable to hear the acoustic difference between synthetic /ba/ and /ga/ as a rising or falling frequency sweep--only an abstract phonetic event /ba/ or /ga/ is heard. It follows that perception of stop consonants is also categorical, i.e., discrimination between physically different acoustic signals is little better than an absolute phonetic categorization (Liberman, Harris, Hoffman, and Griffith, 1957; Mattingly, Liberman, Syrdal, and Halwes, 1971).

In contrast, the perception of nonspeech auditory signals appears "direct" and continuous. When the rising and falling frequency sweeps that differentiate /ba/ and /ga/ are taken out of speech context, listeners hear them in a psychoacoustic manner--one as a rising glissando, the other as a falling glissando. Discrimination between these signals is now equally good or poor both within and between category boundaries (Liberman, Harris, Kinney, and Lane, 1961; Mattingly et al., 1971; Pisoni, 1971).

The perception of speech and nonspeech auditory signals also differs under dichotic competition. Stop consonants are reported better from the right ear (Shankweiler [illegible] Studdert-Kennedy, 1967), presumably because the left hemisphere bears perceptual processing systems specialized for extracting linguistic features from auditory signals (Studdert-Kennedy and Shankweiler, 1970; Liberman, 1973). Nonspeech auditory signals are generally better recognized from the left ear, or show no ear advantage (Kimura, 1964; Chaney and Webster, 1966).

Vowel Perception: One Mode or Two?

Vowels are not encoded to the same degree as stop consonants, especially for a single careful speaker. However, variations in rate of speech for one speaker and differences in vocal tract size between speakers can effectively encode vowels in the speech signal. It is not surprising, then, that vowels have been reported to behave like stop consonants or like nonspeech auditory signals depending on the information-processing constraints imposed by a particular experimental task.

Vowels can be pushed toward continuous or categorical perception as a function of the accessibility of within-category auditory information allowed by the discrimination task (Fujisaki and Kawashima, 1969; Pisoni, 1971). Fujisaki and Kawashima (1969) have found that as vowels are reduced in duration from six glottal pulses to one, discrimination becomes more categorical. Rapidly articulated vowels in word context (Sachs, 1969) and vowels degraded by noise (Lane, 1965) also tend toward categorical perception.

Vowels can also behave either like stop consonants or like nonspeech auditory signals in dichotic listening tasks. Steady-state vowels varying in duration from long to very brief (40 msec) produce no ear advantage (Shankweiler and Studdert-Kennedy, 1967; Darwin, 1969). However, varying vowel formants in relation to vocal tract size (Darwin, 1971a) or emdedding vowels in noise (Weiss and House, 1973) produces the right-ear advantage (REA) expected of stop consonants.

To summarize, many studies suggest that encoded speech signals and nonspeech auditory signals are perceived by means of perceptual systems with necessarily different operating characteristics. Vowels may behave like unencoded auditory signals when there is no demand for the left hemisphere's specialized decoding system. Under certain conditions, however, vowels may engage the specialized phonetic processors of the left hemisphere. The inference we would draw is that when vowels engage these perceptual processing routines, vowel recognition may be disrupted by a following (masking) stimulus. When vowel recognition can be accomplished without engaging the specialized processors, perception is direct and relatively impervious to interference by a following stimulus.

The preceding interpretation of the vowel masking data is supported by the behavior of vowels in a dichotic listening task where the onsets of the stimuli are temporally offset (Porter, Shankweiler, and Liberman, 1969; Studdert-Kennedy, Shankweiler, and Schulman, 1970). In this paradigm better identification of a leading stimulus can be viewed as forward masking; better recognition of a lagging stimulus as backward masking (Darwin, 1971b). Steady-state vowels yield a slight lead effect (Porter et al., 1969). Vowels in CV syllables or in a string of CV syllables show a lag effect or backward masking (Kirstein, 1971).

If backward masking of vowels reflects processing by perceptual mechanisms specialized for speech perception, then other signals which engage the special processor, such as stop consonants (Liberman et al., 1967), should also show backward masking. Indeed, this is the case, as several investigators have reported substantial dichotic and binaural backward masking of stop consonant (Studdert-Kennedy, Shankweiler, and Schulman, 1970; Darwin, 1971b; Porter, 1971; Pisoni, 1972).

The two-processor argument for vowels does not imply the absence of backward recognition masking for nonspeech signals. Backward masking can certainly occur when the perceptual task is made very difficult, e.g., discrimination of stimuli that differ only in harmonic structure (Massaro, 1972a) or discrimination of complex pitch glides (Darwin, 1971b). However, Porter (1971) has reported no backward masking for isolated formant transitions. Evidence for backward recognition masking of pure tones longer than 10 msec duration is equivocal (Massaro, 1972b; Cudahy and Leshowitz, 1974; Leshowitz and Cudahy, in press).

CONCLUSION

The majority of listeners in the present series of experiments evidenced no backward masking even when the target vowels were reduced to minimal duration. This outcome is in marked contrast to the backward masking suffered by stop consonants in a similar paradigm (Pisoni, 1972) and appears to reflect the differences in perceptual processing normally necessary to extract phonetic descriptions of stop consonants and steady-state vowels. Stop consonants, because of their encoded nature, must be processed by a device that can see through the context-conditioned variation in the acoustic signal in order to construct an invariant phonetic message. Construction of the phonetic message takes time, thus making perceptual processing susceptible to interference from a following stimulus. Ordinarily, isolated steady-state vowels do not need the specialized perceptual processing necessary for stop consonants and therefore should, and indeed do, escape backward masking. However, for some listeners 15.5 and 20 msec vowels appear to be so brief as to require a special mode of processing in order to be recognized. For these listeners, increasing the difficulty of extracting target features (Experiment IV) or making target classification more difficult (Experiment III) increased the sensitivity of perceptual processing to interference. Even those listeners who did not appear to use specialized processing to recognize 20 msec vowels resorted to this mode of processing when faced with vowels in noise and, consequently, suffered masking by the second-arriving stimulus.

REFERENCES

Chaney, R. B. and J. Webster. (1966) Information in certain multidimensional sounds. J. Acoust. Soc. Amer. 40, 447-455.

Cooper, F. S. and I. G. Mattingly. (1969) Computer-controlled PCM system for investigation of dichotic speech perception. J. Acoust. Soc. Amer. 46, 115(A).

Cudahy, E. and B. Leshowitz. (1974) Effects of a contralateral interference tone on auditory recognition. Percept. Psychophys. 15, 16-21.

Darwin, C. J. (1969) Auditory perception and cerebral dominance. Unpublished Ph.D. dissertation, University of Cambridge.

Darwin, C. J. (1971a) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 46-62.

Darwin, C. J. (1971b) Dichotic backward masking of complex sounds. Quart. J. Exp. Psychol. 23, 386-392.

Fujisaki, H. and T. Kawashima. (1969) On the codes and mechanisms of speech perception. Annual Report (Division of Electrical Engineering, Engineering Research Institute, University of Tokyo) 1, 67-73.

Haber, R. N. (1969a) Information processing analyses of visual perception: An introduction. In Information Processing Approaches to Visual Perception, ed. by R. N. Haber. (New York: Holt, Rinehart & Winston).

Haber, R. N. (1969b) Repetition, visual persistence, visual noise, and information processing. In Information Processing in the Nervous System, ed. by K. N. Leibovic. (New York: Springer-Verlag).

House, A. (1961) On vowel duration in English. J. Acoust. Soc. Amer. 33, 1174-1178.

Kahneman, D. (1968) Method, findings and theory in studies of visual masking. Psychol. Bull. 70, 404-425.

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exp. Psychol. 16, 355-358.

Kirstein, E. F. (1971) Temporal factors in perception of dichotically presented stop consonants and vowels. Unpublished Ph.D. dissertation, University of Connecticut.

Lane, H. (1965) The motor theory of speech perception: A critical review. Psychol. Rev. 72, 275-309.

Leshowitz, B. and E. Cudahy. (in press) Frequency discrimination in the presence of another tone. J. Acoust. Soc. Amer.

Liberman, A. M. (1973) The specialization of the language hemisphere. In The Neurosciences: Third Study Program, ed. by F. O. Schmitt and F. G. Worden. (Cambridge, Mass.: MIT Press).

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Liberman, A. M., K. S. Harris, J. Kinney, and H. Lane. (1961) The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. 61, 379-388.

Massaro, D. (1972a) Stimulus information vs. processing time and auditory pattern recognition. Percept. Psychophys. 12, 50-57.

Massaro, D. (1972b) Preperceptual auditory images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.

Massaro, D. (1974) Perceptual units in speech recognition. J. Exp. Psychol. 102, 199-208.

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).

Pisoni, D. (1971) On the nature of categorical perception of speech sounds. Ph.D. dissertation, University of Michigan. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Pisoni, D. (1972) Perceptual processing time for consonants and vowels. Haskins Laboratories Status Report on Speech Research SR 31/32, 83-93.

Porter, R. J. (1971) The effect of delayed channel on the perception of dichotically presented speech and nonspeech sounds. Unpublished Ph.D. dissertation, University of Connecticut.

Porter, R. J., D. P. Shankweiler, and A. M. Liberman. (1969) Differential effects of binaural time differences in perception of stop consonants and vowels. J. Acoust. Soc. Amer.

Sachs, R. M. (1969) Vowel identification and discrimination in isolation vs. word context. Quarterly Progress Report (Research Laboratory of Electronics, Massachusetts Institute of Technology) 93, 220-229.

Schiller, P. H. (1965) Monoptic and dichoptic visual masking by patterns and flashes. J. Exp. Psychol. 69, 193-199.

Shankweiler, D. P. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to the left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Sternberg, S. (1966) High speed scanning in human memory. Science 153, 652-654.

Sternberg, S. (1967) Two operations in character recognition: Some evidence from reaction-time measurements. Percept. Psychophys. 2, 45-53.

Studdert-Kennedy, M. (in press) The perception of speech. In Current Trends in Linguistics, ed. by T. A. Sebeok. (The Hague: Mouton).

Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Studdert-Kennedy, M., D. P. Shankweiler, and S. Schulman. (1970) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. Acoust. Soc. Amer. 48, 599-602.

Turvey, M. T. (1973) On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. Psychol. Rev. 80, 1-52.

Weiss, M. and A. House. (1973) Perception of dichotically presented vowels. J. Acoust. Soc. Amer. 53, 51-59.

Vowel and Nasal Durations in Vowel-Nasal-Consonant Sequences in American English: Spectrographic Studies*

Lawrence J. Raphael,[+] Michael F. Dorman,[+] Charles Tobin,[++] and Frances Freeman[+++]

The variation in the duration of vowels caused by the voicing characteristic of the consonants that follow them has been reported for many years in the phonetic/linguistic literature (Kenyon, 1951; Thomas, 1958; House, 1961; Heffner, 1964). The phenomenon is usually described as one in which vowels are lengthened in duration when they precede voiced consonants, so that in such minimal pairs as beat - bead, float - flowed, and bus - buzz the vowel in the first member of each pair is considerably shorter in duration than the vowel in the second member. An approximate estimate of this lengthening effect in English is that the duration of a vowel preceding a voiceless consonant will be from two-thirds to one-half of the duration of that same vowel preceding a voiced consonant (Raphael, 1971). Such an approximation ignores the variance caused by such factors as the intrinsic durations of different vowels (Peterson and Lehiste, 1960) and the manner and place of articulation of the following consonants.

The rather large differences between the durations of English vowels in voiced versus voiceless environments has interested linguists, phoneticians, and speech scientists for a number of reasons--one of the most important being the potential cue value of vowel duration to the perception of a following consonant as either voiced or voiceless.

The existence of this cue value seems likely when one considers that the vowel duration difference may be the only difference consistently present in minimal pairs such as bit - bid, which are distinguished, theoretically, by the voicing or voicelessness of the postvocalic consonant. Those cues to voicing that are operative in other phonetic environments, such as voicing during consonant closure, burst-release, and duration of consonant closure, are often either neutralized or subject to considerable variation in word-final consonants. What this means, of course, if one takes the voiced-voiceless opposition as somehow

---

[*] Portion of a paper presented at the Eighth International Congress on Acoustics, London, July 1974.

[+] Haskins Laboratories, New Haven, Conn., and Herbert H. Lehman College of the City University of New York.

[++] New York University.

[+++] Haskins Laboratories, New Haven, Conn., and the Graduate School and University Center of the City University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

basic or primary to the perceptual separation of minimal pairs such as *bit* - *bid*, is that a secondary feature, vowel duration, is a more significant cue than the supposedly primary cue of voicing. Further, the more significant, though "secondary," feature does not reside within the articulatory period or the acoustic segments being differentiated perceptually, but rather resides within a preceding, adjacent articulatory/acoustic segment.

The cue value of vowel duration has, in fact, been demonstrated (Denes, 1955; Raphael, 1972) for vowels preceding word-final stops, fricatives, and clusters consisting of stop + stop (*lopped* - *lobbed*), stop + fricative (*picks* - *pigs*), and fricative + stop (*bussed* - *buzzed*).

There are, however, several instances in English where vowels are not the sole component of the vocalic segment preceding consonants. These include instances in which a vowel-resonant consonant sequence precedes a voiced or voiceless consonant, yielding such minimal pairs as *bent* - *bend*, *hurt* - *heard*, and *welt* - *weld*. What has been little studied at this point (cf. Lehiste, 1972) and what we have attempted to investigate are the following questions:

1. What are the nature and magnitude of the lengthening effect in vowel + resonant consonant + obstruent consonant syllables?

2. Are the vowel and resonant consonant durations affected differentially by the following consonants?

METHOD AND PROCEDURE

In attempting to determine the magnitude of the lengthening effect for the entire vocalic segment and its component parts, we limited ourselves to spectrographic analyses of monosyllables ending in vowel + nasal + stop sequences. For English, this effectively restricts the utterances studied to those consisting of a vowel + /n/ + /t/ or /d/. Although such sequences as /mp - mb/ (*ample* - *amble*) and /ŋk - ŋg/ (*tinkle* - *tingle*) occur in English, they do not do so in word-final position in monosyllables.

Five native speakers of American English recorded a randomized list of words which included many "dummy" items. The speakers were not told the precise nature of the experiment. The members of the following minimal pairs were dispersed throughout the list: *can't* - *canned*, *daunt* - *dawned*, *stunt* - *stunned*, *pent* - *penned*, *sent* - *send*, *paint* - *pained*, *pint* - *pined*, *mount* - *mound*, and *burnt* - *burned*. Eight different vowels were thus included in the data to be analyzed. Spectrograms were made of each of the utterances and inspected to derive the following data:

1. The duration of the vowel.

2. The duration of the nasal.

3. The duration of the entire vocalic nucleus (vowel + nasal).

The onset of the vowel was taken as that point at which harmonic formant structure first became visible on the spectrogram. Aspiration was thus excluded from the vowel duration. The onset of the nasal consonant was taken as that point at which the weak low-frequency formant, which characterizes nasals in general, was present to the exclusion of the formant structure that characterizes the vowel. That is, in those cases where nasalization accompanied vowel articulation, such nasalization was considered as part of the vowel and not as part of the following nasal consonant.

Although the presence of a nasalized vowel presents certain difficulties to the segmentation of the acoustic signal, we felt secure in our segmenting decisions, more secure than we felt after reading reports of similar research prior to our analysis (Lehiste, 1972).

A more difficult segmentation task occurred at the "boundary" between the nasal consonant and the following stop consonant. The problem here centered around the fact that the nasal formants can be extremely variable over a weak range of intensities. They occasionally can seem to "disappear" for several milliseconds, only to reappear in the middle of what might otherwise have been taken for the closure period of the final stop. Then too, the presence or absence of a fundamental frequency, indicative of voicing, is not of any consistent use in segmenting, since in many cases where the speakers intended voiceless /t/, vocal fold vibrations occur well into the stop closure period, and since, in the opposing case, intended voiced /d/ revealed little, if any, voicing during the occlusion for the final stop. This apparent inconsistency of glottal pulsing during closure is to be expected if it is in fact the duration of some part or parts of the preceding vocalic element which cues the perception of voicing in the final consonant. That is, the feature that embodies the more important perceptual cue is preserved; the feature bearing a redundant, less important perceptual cue is subject to deletion--or perhaps "atrophy" would be a more appropriate word.

The most satisfactory solution for segmenting the nasal from the final stop seemed to be to mark the termination of the nasal at the last point in time at which one could confidently discern the low-frequency nasal formant, if a fundamental frequency was also present.

RESULTS

Two major findings resulted from the spectrographic analyses. First, both the vowel and the nasal consonant, and thus the entire preceding vocalic segment, were of greater duration when they preceded /d/ than when they preceded /t/. This was true without exception for all speakers and all utterances (Table 1). Although there were instances where the magnitude of the increase in duration was quite small for either the nasal or the vowel in an utterance (as little as 5 msec for the vowel and 15 msec for the nasal), the smallest durational difference found between any single vocalic nucleus as it occurred in each voicing environment was 55 msec. It appears that if one of the components in the vocalic nucleus increases only minimally in duration before a voiced consonant, then the other component will undergo a more substantial durational increment.

Second, the vowel and nasal durations were affected differentially by their voicing environments (Tables 2 and 3). Although both vowel and nasal increased in duration from the voiceless to the voiced environments, the increment of nasal duration was proportionately greater than that of vowel duration. For example, for speaker #1 (Table 2) it can be seen that in the word pent the vowel represents 68.3% of the vocalic nucleus, the nasal 31.7%. But in the word pend, the vocalic nucleus is evenly divided between vowel and nasal. Table 3 reveals, for

TABLE 1: Vowel, nasal, and vowel + nasal durations for five speakers. Values (in msec) are rounded to the nearest 5 msec.

	Speaker #1			Speaker #2			Speaker #3			Speaker #4			Speaker #5		
	V	/n/	V+/n/	V	/n/	V+/n/	V	/n/	V+/n/	V	/n/	V+/n/	V	/n/	V+/n/
PENT	150	70	220	80	75	155	130	75	205	90	55	145	85	40	125
PEND	175	175	350	155	255	410	140	125	265	140	160	300	110	95	205
SENT	145	65	210	100	65	165	130	70	200	105	70	175	70	65	135
SEND	180	130	310	110	195	305	180	110	290	155	110	265	80	115	195
PAINT	135	65	200	110	40	150	145	40	185	140	35	175			
PAINED	255	135	390	160	210	370	175	90	265	205	95	300			
CAN'T	175	60	235	150	75	225	170	45	215	175	30	205	125	45	170
CANNED	225	140	365	220	170	390	255	70	325	215	110	325	150	75	225
DAUNT	235	60	295	160	90	255	225	75	300	220	30	250	160	30	190
DAWNED	260	125	385	215	230	445	255	125	380	260	90	350	200	45	245
STUNT	145	35	180	75	80	155	105	60	165	110	55	165	100	25	125
STUNNED	150	160	310	100	175	275	150	100	250	145	145	290	110	80	190
PINT	175	60	235	110	45	155	195	40	235	160	55	215			
PINED	245	115	360	205	170	375	245	60	305	300	105	405			
MOUNT	170	80	250				200	40	240	205	50	255			
MOUND	295	130	425				280	60	340	265	95	360			
BURNT				100	45	145	135	50	185	155	75	230	110	25	135
BURNED				165	135	300	210	75	285	210	145	355	195	75	270

TABLE 2: Percentages of vocalic nucleus distributed between vowel and nasal.

	Speaker #1		Speaker #2		Speaker #3		Speaker #4		Speaker #5	
	V	/n/	V	/n/	V	/n/	V	/n/	V	/n/
PENT	68.3	31.7	51.4	48.6	63.5	36.5	62.0	38.0	68.1	32.9
PEND	50.0	50.0	37.8	62.2	52.9	47.1	46.7	53.3	53.8	46.2
SENT	69.2	30.8	60.7	39.3	65.0	35.0	60.1	39.9	51.4	48.6
SEND	58.1	41.9	36.2	63.8	62.1	37.9	58.6	41.4	40.9	59.1
PAINT	67.5	32.5	73.5	26.5	78.5	21.5	80.0	20.0		
PAINED	65.4	34.6	31.8	68.2	66.1	33.9	74.2	26.8		
CAN'T	74.5	25.5	66.8	33.2	79.0	21.0	85.5	14.5	73.7	26.3
CANNED	61.6	39.4	56.4	43.6	78.5	21.5	66.2	33.8	66.8	33.2
DAUNT	79.8	20.2	62.8	37.2	75.0	25.0	88.0	12.0	84.2	16.8
DAWNED	67.5	32.5	48.3	51.7	68.9	31.1	74.4	25.6	81.6	18.4
STUNT	51.9	48.1	48.6	51.4	63.7	36.3	66.7	33.3	80.0	20.0
STUNNED	48.4	51.6	36.4	63.6	60.0	40.0	50.0	50.0	55.3	44.7
PINT	74.5	25.5	71.0	29.0	83.0	17.0	74.4	25.6		
PINED	68.0	32.0	54.7	45.3	80.5	19.5	74.1	25.9		
MOUNT	68.0	32.0			83.5	16.5	80.5	19.5		
MOUND	69.5	30.5			82.4	17.6	73.7	26.3		
BURNT			69.0	31.0	73.0	27.0	67.5	32.5	81.6	18.4
BURNED			55.0	45.0	73.8	26.2	59.2	40.8	72.3	27.7

TABLE 3: Percentages of increase for vowel, nasal, and vowel + nasal from voiceless to voiced environments.

	Speaker #1			Speaker #2			Speaker #3			Speaker #4			Speaker #5		
	V	/n/	V+/n/	V	/n/	V+/n/	V	/n/	V+/n/	V	/n/	V+/n/	V	/n/	V+/n/
PENT PEND	16.7	150.0	59.1	90.5	230.0	158.5	7.7	66.7	29.2	55.5	191.1	106.8	29.6	112.2	64.0
SENT SEND	24.1	100.0	47.6	10.0	207.7	84.8	38.4	57.1	45.0	47.6	57.2	51.4	14.3	77.0	44.4
PAINT PAINED	88.9	107.7	95.0	45.4	425.0	146.6	26.5	124.9	43.2	46.4	171.5	41.7			
CAN'T CANNED	28.5	133.3	55.3	46.4	120.0	74.1	50.0	66.7	51.2	22.8	286.0	58.6	20.0	66.7	32.4
DAUNT DAWNED	10.6	108.3	30.5	33.3	150.0	74.5	13.3	66.7	26.6	18.2	200.0	40.0	25.0	50.0	28.9
STUNT STUNNED	3.4	357.1	72.3	31.6	114.3	77.4	42.8	66.7	51.5	31.8	163.8	75.8	10.0	22.0	52.0
PINT PINED	40.0	91.7	53.2	86.4	277.8	141.9	25.6	50.0	29.8	87.6	91.1	88.6			
MOUNT MOUND	73.5	62.5	70.0				40.0	50.0	41.7	29.3	90.0	41.2			
BURNT BURNED				65.0	200.0	106.9	56.5	50.0	54.1	35.5	93.4	54.4	77.2	200.0	100.0

the same subject and utterances, that the vowel in *pend* has increased by 16.7% of its duration in *pent*, whereas the nasal has increased by 150.0% of its duration. Nor is this by any means an extreme case, as inspection of the data in the tables will reveal.

A final observation concerns the two cases in which a reversal occurs in the data. The pairs *mount* - *mound* (speaker #1) and *burnt* - *burned* (speaker #3) reveal that the vowel duration increase is proportionately greater than the increase in nasal consonant duration. Inspection of these two cases reveals an extensive nasalization of the vowel in each. It may well be that an alternative strategy to a proportionately greater lengthening of the nasal duration is the addition of nasal resonances to the vowel. Further investigations are needed to determine if there is a regular trading relationship between the duration of vowel nasalization and the duration of the nasal per se.

DISCUSSION

The results of these spectrographic analyses suggest two hypotheses. First, that both the individual vowel and nasal durations, as well as their combined duration (i.e., that of the vocalic nucleus) are potentially sufficient cues to the voicing characteristic of the following consonant. Second, that the proportionately greater variation in nasal duration indicates that it should be a more powerful cue than that of vowel duration to the voicing characteristic of the following consonant. Perceptual experiments designed to test both these hypotheses are planned for the near future.

Finally, although it is reasonable to assume that these results for vowel-nasal-stop sequences may be generalized to sequences with other types of resonant consonants (e.g., /l/), and other types of final consonants (e.g., fricatives), the data for these latter types of sequences should be gathered and, if warranted, followed by the appropriate perceptual experiments.

REFERENCES

Denes, P. (1955) Effect of duration on the perception of voicing. J. Acoust. Soc. Amer. *27*, 761-764.

Heffner, R-M. S. (1964) *General Phonetics*. (Madison, Wis.: University of Wisconsin Press).

House, A. S. (1961) On vowel duration in English. J. Acoust. Soc. Amer. *33*, 1174-1178.

Kenyon, J. S. (1951) *American Pronunciation*, 10th ed. (Ann Arbor, Mich.: George Wahr).

Lehiste, I. (1972) Manner of articulation, parallel processing, and the perception of duration. Working Papers in Linguistics (Computer and Information Science Research Center, Ohio State University) *12*, 33-52.

Peterson, G. E. and I. Lehiste. (1960) Duration of syllabic nuclei in English. J. Acoust. Soc. Amer. *32*, 693-703.

Raphael, L. J. (1971) Vowel duration as a cue to the perceptual separation of cognate sounds in American English. Ph.D. dissertation, City University of New York. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. J. Acoust. Soc. Amer. *51*, 1296-1303.

Thomas, C. K. (1958) *An Introduction to the Phonetics of American English*, 2nd ed. (New York: Ronald Press).

Vowel and Nasal Durations as Perceptual Cues to Voicing in Word-Final Stop Consonants*

Michael F. Dorman,[+] Lawrence J. Raphael,[+] Frances Freeman,[++] and Charles Tobin[+++]

A recent spectrographic study of the variation in duration of the component segments of vowel-plus-nasal sequences preceding voiced or voiceless alveolar stop consonants (Raphael, Dorman, Tobin, and Freeman, 1974) revealed the following:

1. Both the vowel and the nasal consonant (and thus the entire vocalic nucleus) are of greater duration before /d/ than before /t/.

2. If the voiceless case is assumed as a base, then the increase in nasal duration in the voiced case is proportionately greater than the increase in vowel duration.

These findings suggest that while vowel duration may provide some part of the perceptual cue to voicing, as it does in CVC syllables (Denes, 1955; Raphael, 1972), the proportionately greater variation in nasal duration may provide a more powerful cue. To test this hypothesis, listeners were presented with synthetic CVnC utterances, in which the nasal and vowel segments were independently varied, and were asked to label the final consonant as voiced /d/ or voiceless /t/.

METHOD

Subjects

The listeners were ten student volunteers from Herbert H. Lehman College.

Preparation of Stimuli

Two series of stimuli were prepared on the Haskins Laboratories' parallel resonance synthesizer. The stimuli in each series were variations of bend (/bɛnd/), synthesized as follows:

---

*Portion of a paper presented at the Eighth International Congress on Acoustics, London, July 1974.

[+] Haskins Laboratories, New Haven, Conn., and Herbert H. Lehman College of the City University of New York.

[++] Haskins Laboratories, New Haven, Conn., and the Graduate School and University Center of the City University of New York.

[+++] New York University.

[HASKINS LABORATORIES: Status Report on Speech Research SR-37/38 (1974)]

1. Thirty-five msec transitions to a three-formant /ɛ/, appropriate to both stop manner and bilabial place of articulation. A F_o of 100 Hz extended throughout the duration of all stimuli.

2. A steady-state vowel, /ɛ/, with F_1 = 537 Hz, F_2 = 1845 Hz, and F_3 = 2525 Hz, followed by linear formant transitions to the nasal consonant /n/.

3. A steady-state nasal consonant, /n/, with weak F_1 and F_2 at 260 Hz and 1232 Hz, respectively, and with F_3 at 2358 Hz.

4. A formantless (closure) interval of 100 msec.

5. A three-formant, 10-msec, frictionless burst-release, with formant frequencies appropriate to stop manner and alveolar place of articulation (see Figure 1).

In the first series of stimuli the nasal duration was held constant at 130 msec while the vowel duration varied between 40 msec and 200 msec in 20-msec steps. In the second stimulus series the nasal duration varied in 20-msec steps over a range of 40 msec to 200 msec, while the vowel duration was held constant at 130 msec. Spectrographic evidence reported in a previous experiment (Raphael et al., 1974) indicated that the 130 msec constant duration for both vowel and nasal was representative of real-speech values.

A separate test tape was prepared for each of the two stimulus series. Each tape contained 6 tokens of each stimulus type, for a total of 54 test items. The stimuli were randomized on the tape. The interstimulus interval was 3 sec.

Design

Two groups of listeners were tested separately. One group of listeners (n=5) heard the nasal-varying series first, then the vowel-varying series. A second group (n=5) heard the vowel-varying sequence first, then the nasal-varying sequence.

Procedure

The listeners were seated in a semicircle, within a large, sound-attenuated room, facing an AR-4x loudspeaker. The listeners were told they would hear a series of computer-synthesized words, and were instructed to label on a printed response form the final consonant as either /d/ or /t/. After several tokens of the stimuli were presented to familiarize the listeners with the synthetic signals, the first test sequence was begun. At the end of this sequence, the listeners were given a brief rest, then were presented with the second test sequence.

RESULTS

Vowel Duration Series

All of the subjects indicated a change in perception of the final stop consonant from the short to the long end of the vowel duration range. When the vowel was at the short end of the range, all listeners reported the stimuli as ending in /t/; when the vowel duration was long, all listeners reported /d/ as the final consonant.

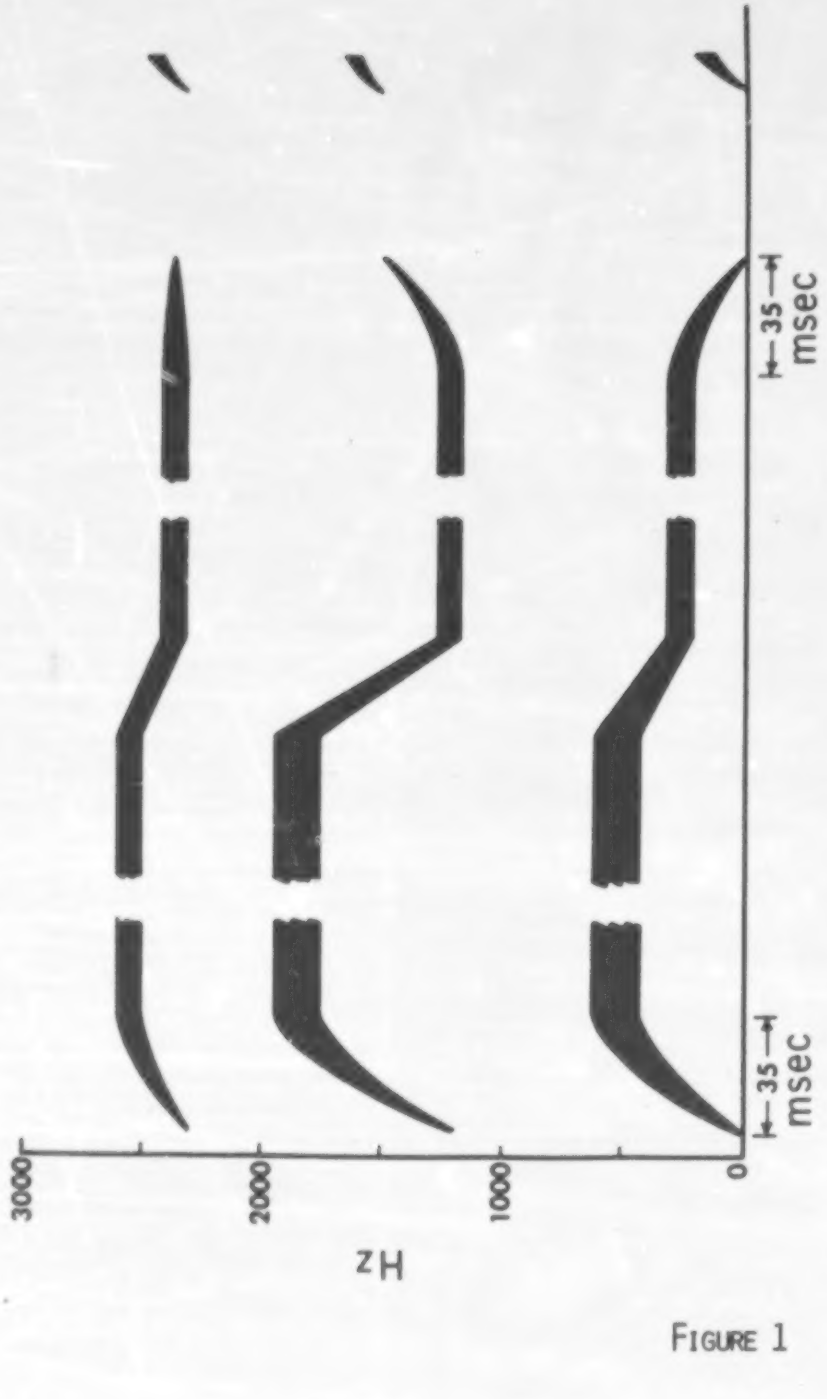


FIGURE 1

The 50% point in listeners' judgments occurred over a range of vowel durations of from 115 msec to 180 msec (see Figure 2). The mean crossover point for all listeners was 142.5 msec.

Nasal Duration Series

Again, all listeners indicated a change in perception of the final consonant. Short duration nasals elicited /t/ identifications, whereas long duration nasals elicited /d/ as a response. The 50% point in listeners' judgments occurred over a range of 90 msec to 115 msec (see Figure 3). The mean crossover point for all listeners was 99.4 msec of nasal duration. Figure 4 displays the ranges and means of the crossover points for the vowel and nasal series of stimuli.

DISCUSSION

The perceptual data appear to support the hypothesis that the word-final voiced-voiceless opposition, at least as represented by /t/ and /d/ in this experiment, can be cued by the duration of either the preceding vowel or the nasal when in a vowel-nasal-stop sequence. By extension, therefore, the duration of the entire vocalic segment of vowel + nasal should be a cue to the perception of voicing in a following consonant.

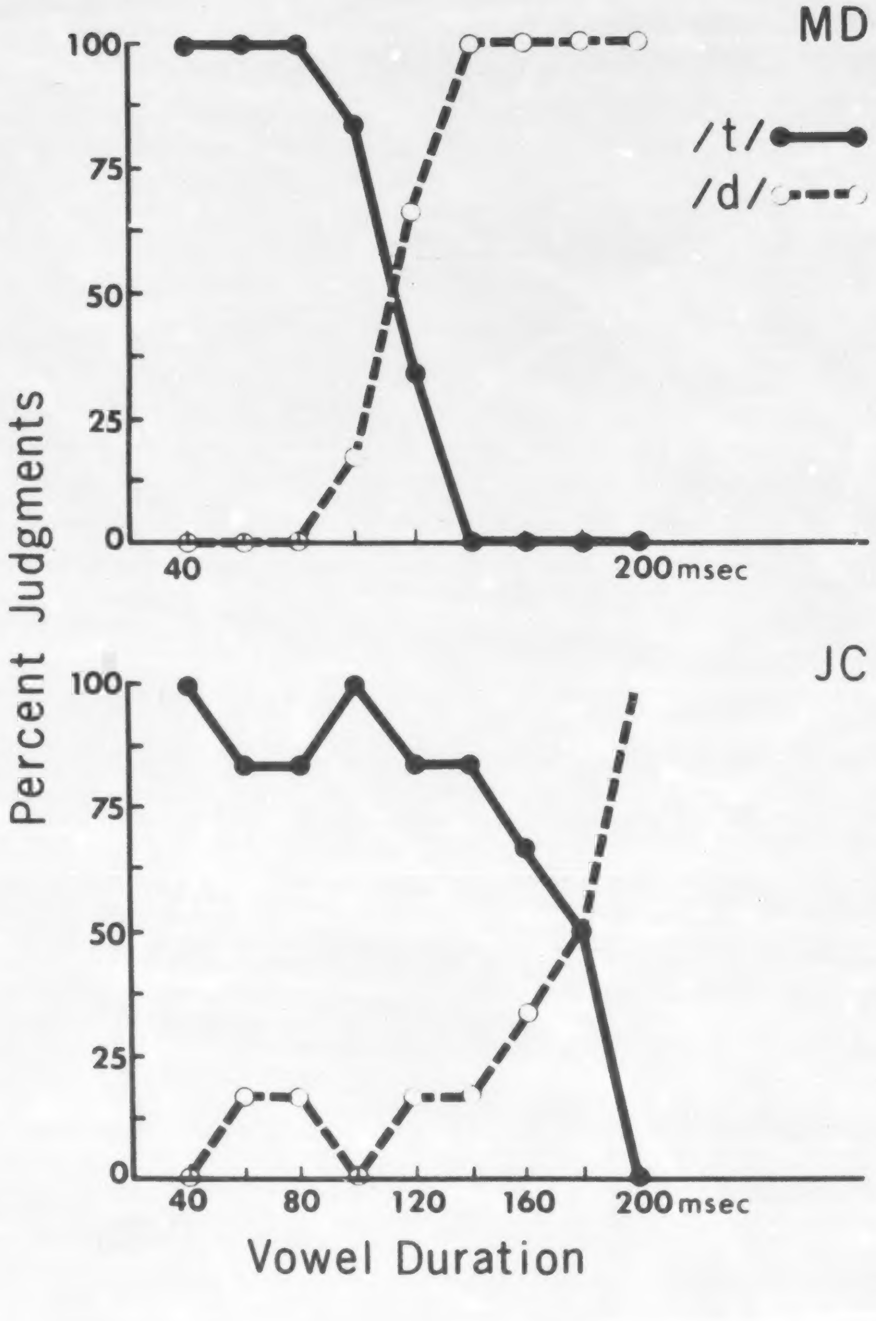
The data also reveal that the duration of the nasal is a relatively stronger cue than that of vowel duration. This seems clear from a comparison of the amount of variation in nasal versus vowel duration needed to change the listeners' judgments from /t/ to /d/. Clearly, listeners' perceptions of final consonants as voiced or voiceless are more sensitive to variations in nasal duration than to variations in vowel duration. Assuming that the cues to voicing contained in the final consonant segment do have some effect on perception, and recalling that these cues in the synthetic stimuli were appropriate to /d/, it is apparent that the effect of these cues can be either neutralized or enhanced by smaller changes in nasal duration than in vowel duration.

Such an outcome seems reasonable, if for no other reason than the greater proximity of the nasal than of the vowel to the following consonant. If the durations of the vowel and nasal segments are stored separately in echoic memory during speech perception, then it may simply be the case that the shorter storage time needed for the nasal segment makes it more efficient as a primary cue. The cue value of the more distant (vowel) segment may be reduced (1) by the loss of some of the auditory information from echoic memory during the processing of the closer (nasal) segment, (2) by interference from the entry of the nasal information into echoic memory, or (3) by a combination of (1) and (2). Any of these mechanisms could also contribute to the fact that although the cue of vowel duration is relatively weaker, it can still be a sufficient cue to the perception of the voicing characteristic of the word-final consonant.

REFERENCES

Denes, P. (1955) Effect of duration on the perception of voicing. J. Acoust. Soc. Amer. 27, 761-764.

Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. J. Acoust. Soc. Amer. 51, 1296-1303.



MD
/t/
/d/
100
75
50
25
0
40
200 msec
Percent Judgments
JC
100
75
50
25
0
40
80
120
160
200 msec
Vowel Duration

FIGURE 2

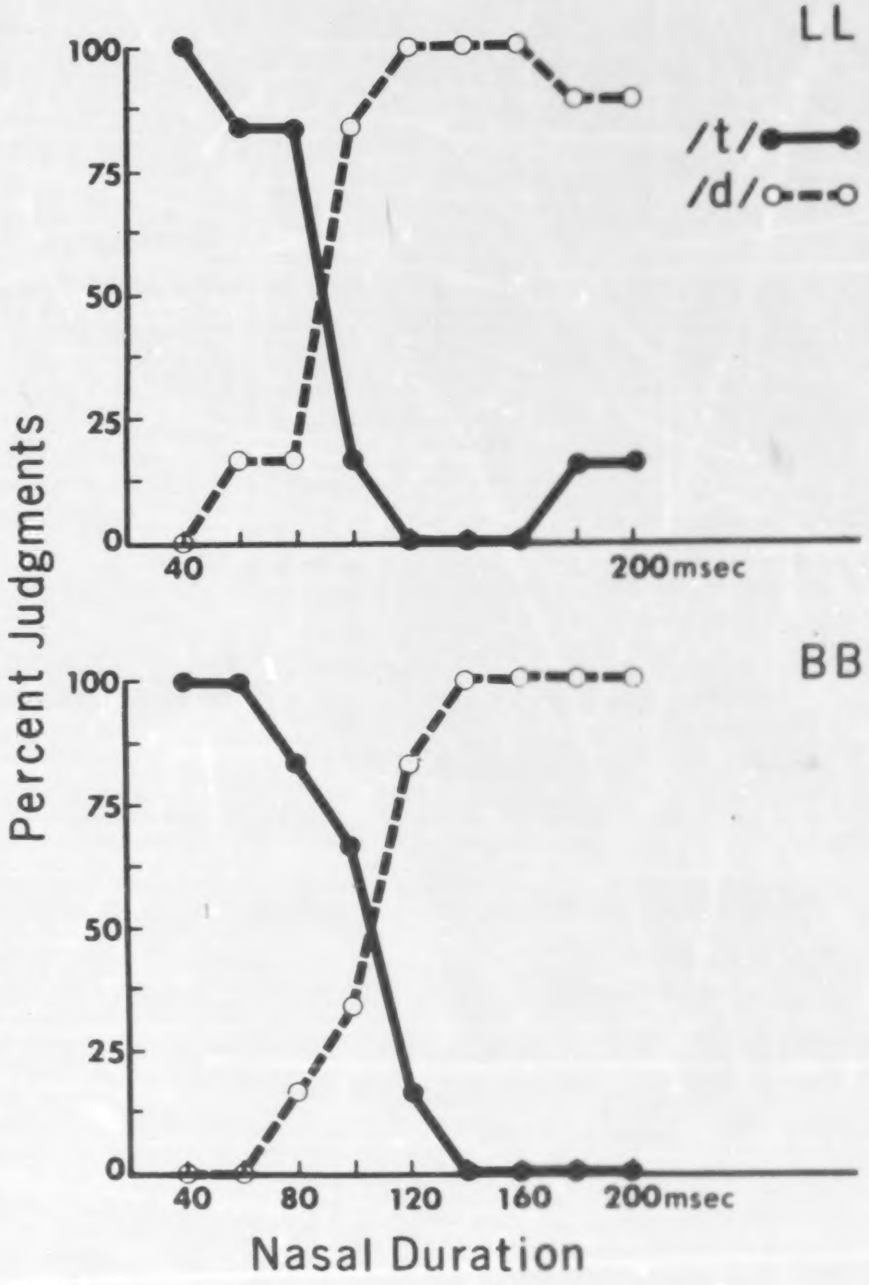

LL
/t/
/d/
100
75
50
25
0
40
200msec
Percent Judgments
BB
100
75
50
25
0
40
80
120
160
200msec
Nasal Duration

FIGURE 3

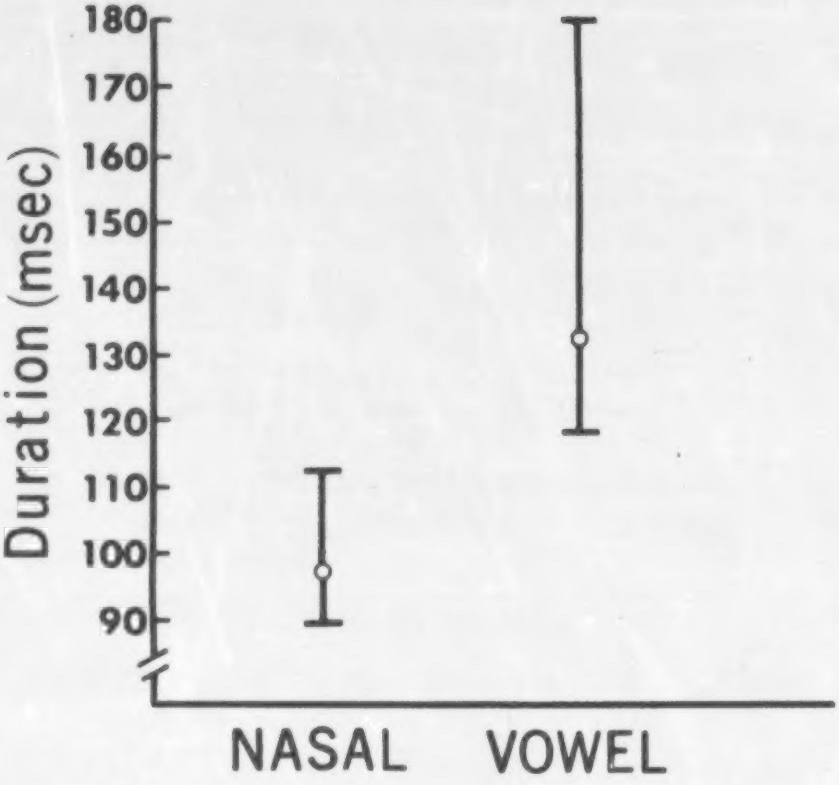

Duration (msec)
180
170
160
150
140
130
120
110
100
90
NASAL
VOWEL

FIGURE 4

Raphael, L. J., M. F. Dorman, C. Tobin, and F. Freeman. (1974) Vowel and nasal durations in vowel-nasal-consonant sequences in American English: Spectrographic studies. Haskins Laboratories Status Report on Speech Research SR-37/38 (this issue).

III. PUBLICATIONS AND REPORTS

IV. APPENDIX

PUBLICATIONS AND REPORTS

Publications and Manuscripts

The following two papers were reprinted in Human Communication: A Unified View, ed. by Edward E. David and Peter B. Denes (New York: McGraw Hill, 1972):

Perception of the Speech Code. A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. 13-50.

Speaker Identification by Speech Spectrograms: A Scientist's View of its Reliability for Legal Purposes. Richard H. Bolt, Franklin S. Cooper, Edward E. David, Peter B. Denes, James M. Pickett, and Kenneth N. Stevens. 369-398.

The Phi Coefficient as an Index of Ear Differences in Dichotic Listening. G. M. Kuhn. Cortex (1973) 9, 450-457.

The Development of Auditory Feedback Monitoring: Delayed Auditory Feedback Studies on the Vocalizations of Children Aged Six Months to 19 Months. N. Fargo Belmore, Diane Kewley-Port, Richard L. Mobley, and Violet E. Goodman. Journal of Speech and Hearing Research (1973) 16, 700-708.

Short-Term Habituation of the Infant Auditory Evoked Response. Michael F. Dorman and Robert Hoffmann. Journal of Speech and Hearing Research (1973) 16, 637-641.

Auditory and Linguistic Processes in the Perception of Intonation Contours. Michael Studdert-Kennedy and Kerstin Hadding. Language and Speech (1973) 16, 293-313.

On the Evolution of Human Language. Philip Lieberman. In A Festschrift for Morris Halle, ed. by Stephen Anderson and Paul Kiparsky. (New York: Holt, Rinehart & Winston, 1973).

Oral Feedback I: Variability of the Effect of Nerve-Block Anesthesia Upon Speech. Gloria Jones Borden, Katherine S. Harris, and William Oliver. Journal of Phonetics (1973) 1, 289-295.

Oral Feedback II: An Electromyographic Study of Speech Under Nerve-Block Anesthesia. Gloria Jones Borden, Katherine S. Harris, and Lorne Catena. Journal of Phonetics (1973) 1, 297-308.

Auditory Evoked Potential Correlates of Speech Sound Discrimination. Michael F. Dorman. Perception and Psychophysics (1974) 15, 215-220.

Dichotic Release from Masking for Speech. Timothy C. Rand. Journal of the Acoustical Society of America (1974) 55, 678-680.

Are You Asking Me, Telling Me, or Talking to Yourself? Kerstin Hadding and Michael Studdert-Kennedy. Journal of Phonetics (1974) 2, 7-14.

Effect of Speaking Rate on Labial Consonant-Vowel Articulation. Thomas Gay, Tatsujiro Ushijima, Hajime Hirose, and Franklin S. Cooper. *Journal of Phonetics* (1974) *2*, 47-63.

Laryngeal Control in Korean Stop Production. H. Hirose, C. Y. Lee, and T. Ushijima. *Journal of Phonetics* (1974) *2*, 145-152.

The Specialization of the Language Hemisphere. A. M. Liberman. In *The Neurosciences: Third Study Program*, ed. by Francis O. Schmitt and Frederick G. Worden. (Cambridge, Mass.: MIT Press, 1974) 43-56.

*A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech. Paul Mermelstein. (Presented at the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Pa., 15-19 April 1974.) In *IEEE Symposium on Speech Recognition: Contributed Papers*. (New York: IEEE, 1974) 144-147.

The Function of Features. Michael Studdert-Kennedy. Paper presented at the Mount Sinai Conference on Language Disorders, 21 March 1974; to be published in *Journal of Communication Disorders*.

The following five papers, presented at a National Institute of Child Health and Human Development conference, "The Role of Speech in Language," held at Columbia, Md., October 1973, will be published in the conference proceedings, ed. by J. F. Kavanagh and J. E. Cutting (Cambridge, Mass.: MIT Press):

> *The Role of Speech in Language: Introduction to the Conference. Alvin M. Liberman
>
> *The Human Aspect of Speech. Ignatius G. Mattingly
>
> *From Continuous Signal to Discrete Message: Syllable to Phoneme. Michael Studdert-Kennedy
>
> *The Evolution of Speech and Language. Philip Lieberman
>
> *Phonetic Feature Analyzers and the Processing of Speech in Infants. James E. Cutting and Peter D. Eimas

The Function of Phonetic Categories. Michael Studdert-Kennedy. Paper presented at the Indiana Theoretical and Cognitive Psychology Conference, Bloomington, Ind., 4 April 1974; to be published in *Cognitive Theory*, ed. by F. Restle. (Potomac, Md.: Erlbaum Press).

The Intonation of Verifiability. R. Nash and A. Mulac. In *Prosodica: Papers in Honor of Dwight L. Bolinger*, ed. by L. Waugh and C. H. Van Schooneveld. (The Hague: Mouton, in press).

---

*Appears in this report, SR-37/38.

Reaction Times to Comparisons Within and Across Phonetic Categories: Evidence for Auditory and Phonetic Levels of Processing. David B. Pisoni and Jeffrey Tash. Perception and Psychophysics (in press). (Also in SR-34, 77-88.)

On the Identification of Place and Voicing Features in Synthetic Stop Consonants. James R. Sawusch and David B. Pisoni. Journal of Phonetics (in press). (Also in SR-35/36, 65-80.)

Auditory Short-Term Memory and Vowel Perception. David B. Pisoni. Memory and Cognition (in press). (Revised version of SR-34, 89-117.)

On the Short-Term Retention of Serial, Tactile Stimuli. Edie V. Sullivan and M. T. Turvey. Memory and Cognition (in press). (Also in SR-33, 123-135.)

Early Apical Stop Production: A Voice-Onset Time Analysis. Diane Kewley-Port and Malcolm S. Preston. Journal of Phonetics (in press).

Speech Perception. Michael Studdert-Kennedy. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (Springfield, Ill.: C. C Thomas, in press).

*Different Speech-Processing Mechanisms Can be Reflected in the Results of Discrimination and Dichotic Listening Tasks. James E. Cutting. To appear in Brain and Language.

*Identification of Vowel Order: Concatenated Versus Formant-Connected Sequences. M. F. Dorman, James E. Cutting, and Lawrence J. Raphael.

*Hemispheric Lateralization for Speech Perception in Stutterers. M. F. Dorman and R. J. Porter, Jr.

*Dichotic Release from Masking: Further Results from Studies with Synthetic Speech Stimuli. P. W. Nye, T. M. Nearey, and T. C. Rand.

*Categories and Boundaries in Speech and Music. James E. Cutting and Burton S. Rosner.

*The Intelligibility of Synthetic Monosyllabic Words in Short, Syntactically Normal Sentences. P. W. Nye and J. H. Gaitenby.

*An Experimental Evaluation of the EMG Data Processing System: Time Constant Choice for Digital Integration. Diane Kewley-Port.

Reports and Oral Presentations

Stages Underlying Perception at a Glance. M. T. Turvey. Presented at University of Sussex, September 1973; Oxford University, October 1973; Birkbeck College and University College, University of London, October 1973; University of Reading, October 1973; Queens University, Belfast, November 1973; and Sheffield University, November 1973.

The Direct Viewing of Velar Movements During Speech. Tatsujiro Ushijima and H. Hirose. Presented at the American Speech and Hearing Association meeting, Detroit, Mich., 14 October 1973.

Language, Speech, and Brain. Michael Studdert-Kennedy. New York Academy of Sciences, 30 October 1973.

Action and Perception. M. T. Turvey. Invited address to the British Psychological Association (Northern Ireland), Belfast, November 1973.

Hemispheric Specialization for Speech Perception. Michael Studdert-Kennedy. Queens University, Kingston, Ontario, 8 November 1973.

*On "Explaining" Vowel Duration Variation. Leigh Lisker. Presented at the winter meeting of the Linguistic Society of America, San Diego, Calif., 28-30 December 1973.

Computer Simulation of the Human Speech Production System. Paul Mermelstein. Engineering Colloquium, University of Bridgeport, Conn., 5 December 1973.

New Developments in Brain Function for Speech Perception and Production, Symposium. Ruth S. Day. American Academy for the Advancement of Science. San Francisco, Calif., 28 February 1974.

Rivalry, Fusion, and "Cyclotean" Perception. James E. Cutting. Colloquia given at University of Pennsylvania, Philadelphia; Princeton University, Princeton, N. J.; Wesleyan University, Middletown, Conn.; Cornell University, Ithaca, N. Y.; and University of Connecticut, Storrs, February 1974.

The Acoustic Syllable. Michael Studdert-Kennedy. Department of Psychology, Massachusetts Institute of Technology, Cambridge, Mass., 22 February 1974.

Physiological Linguistics: The State of the Art. Lawrence J. Raphael. Presented at the annual conference of the International Linguistic Association, New York, March 1974.

Velopharyngeal Closure in Normal Speakers. Katherine S. Harris. New York State Speech and Hearing Association, South Fallsburg, N. Y., 29 April 1974.

Assessing the Intelligibility of a Prototype Reading Machine for the Blind. P. W. Nye and J. H. Gaitenby. Presented at the annual convention of the International Communications Association, New Orleans, La., 17-20 April 1974.

The following papers were presented at the 87th meeting of the Acoustical Society of America, New York, April 1974.

*What Information Enables a Listener to Map a Talker's Vowel Space? Robert Verbrugge, Winifred Strange, and Donald Shankweiler

*Consonant Environment Specifies Vowel Identity. Winifred Strange, Robert Verbrugge, and Donald Shankweiler

*Binaural Subjective Tones and Melodies Without Monaural Familiarity Cues. Michael Kubovy, James E. Cutting, and Roderick McI. McGuire

*The Function of the Posterior Cricoarytenoid in Speech Articulation. Hajime Hirose and Tatsujiro Ushijima

*More on the Motor Organization of Speech Gestures. Fredericka Bell-Berti and Katherine S. Harris

*Laryngeal Activity Accompanying the Moment of Stuttering: A Preliminary Report of EMG Investigations. Frances J. Freeman and Tatsujiro Ushijima

*Two Processes in Vowel Recognition: Inferences from Studies of Backward Masking. M. F. Dorman, D. Kewley-Port, S. Brady, and M. T. Turvey

*Electromyographic Study of the Velum During Speech. T. Ushijima and H. Hirose

Word Recall in Aphasia. Diane Kewley-Port

Segmentation of Speech into Syllabic Units. Paul Mermelstein and G. M. Kuhn

Category Boundaries for Linguistic and Nonlinguistic Dimensions of the Same Stimuli. J. R. Sawusch, D. B. Pisoni, and J. E. Cutting

Colloquia. Ruth S. Day. Oxford University, Oxford, England, 9 January 1974; Vanderbilt University, Nashville, Tenn., 24 January 1974; University of California, Los Angeles (Perception and Psycholinguistics), 25 February 1974; California Institute of Technology, Division of Biology, Pasadena, Calif., 26 February 1974; University of Michigan, Mental Health Research Institute, Ann Arbor, 9 May 1974; and California Institute of Technology, Pasadena, Calif., 20 May 1974.

Invited Addresses. Ruth S. Day. Experimental Psychology Society, London, England, 4 January 1974; Learning and Research Development Center Conference on the Nature of Intelligence, Pittsburgh, Pa., 5 March 1974; and Social Science Research Council Conference (Modes of Perceiving), University of Minnesota, Minneapolis, 27 June 1974.

An Introduction to Recent Topics in Speech Science, Panel Discussion. New Jersey Speech and Hearing Association, Great Gorge, N. J., 3 May 1974:

Research Techniques. F. Bell-Berti

Questions We Can Ask About Speech Production. Katherine S. Harris

Some Answers Related to Diagnosis and Therapy. Gloria G. Borden

Phonetic Recoding and the Beginning Reader. I. Y. Liberman and D. P. Shankweiler. Invited paper, Basic Research on the Reading Process Conference, City University of New York, Graduate Center, 18 May 1974.

The following two papers will be presented at the Eighth International Congress on Acoustics, London, July 1974, under the title "Vowel and Nasal Duration as Cues to the Perceptual Categorization of Word-Final Consonants in English":

*Vowel and Nasal Durations in Vowel-Nasal-Consonant Sequences in American English: Spectrographic Studies. Lawrence J. Raphael, Michael F. Dorman, Charles Tobin, and Frances Freeman.

*Vowel and Nasal Durations as Perceptual Cues to Voicing in Word-Final Stop Consonants. Michael F. Dorman, Lawrence J. Raphael, Frances Freeman, and Charles Tobin.

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers:

SR-21/22 to SR-35/36

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October- December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	

AD numbers may be ordered from: U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from: ERIC Document Reproduction Service
Leasco Information Products, Inc.
P. O. Drawer O
Bethesda, Maryland 20014